Covariance Estimation in High Dimensions via Kronecker Product Expansions

Theodoros Tsiligkaridis *, *Student Member, IEEE*, Alfred O. Hero III, *Fellow, IEEE*

Abstract

This paper presents a new method for estimating high dimensional covariance matrices. Our method, permuted rank-penalized least-squares (PRLS), is based on a Kronecker product series expansion of the true covariance matrix. Assuming an i.i.d. Gaussian random sample, we establish high dimensional rates of convergence to the true covariance as both the number of samples and the number of variables go to infinity. For covariance matrices of low separation rank, our results establish that PRLS has significantly faster convergence than the standard sample covariance matrix (SCM) estimator. In addition, this framework captures a fundamental tradeoff between estimation error and approximation error, thus providing a scalable covariance estimation framework in terms of separation rank, an analog to low rank approximation of covariance matrices [1]. The MSE convergence rates generalize the high dimensional rates recently obtained for the ML Flip-flop algorithm [2], [3] for Kronecker product covariance estimation. We show that a class of block Toeplitz covariance matrices has low separation rank and give bounds on the minimal separation rank r that ensures a given level of bias. Simulations are presented to validate the theoretical bounds. As a real world application, we illustrate the utility of the proposed Kronecker covariance estimator in spatio-temporal linear least squares prediction of multivariate wind speed measurements.

Index Terms

Structured covariance estimation, penalized least squares, Kronecker product decompositions, high dimensional convergence rates, mean-square error, multivariate prediction.

I. Introduction

Covariance estimation is a fundamental problem in multivariate statistical analysis. It has received attention in diverse fields including economics and financial time series analysis (e.g., portfolio selection, risk management and asset pricing [4]), bioinformatics (e.g. gene microarray data [5], [6], functional MRI [7]) and machine learning (e.g., face recognition [8], recommendation systems [9]). In many modern

applications, data sets are very large with both large number of samples n and dimension d, often with $d \gg n$, leading to a number of covariance parameters that greatly exceeds the number of observations. The search for good low-dimensional representations of these data sets has recently led to breakthroughs in multivariate statistics and signal processing. Recent examples include sparse covariance estimation [10], [11], [12], [13], low rank covariance estimation [14], [15], [16], [1], and Kronecker product esimation [17], [18], [19], [2], [3].

Kronecker product (KP) structure is a different covariance constraint from sparse or low rank constraints. KP represents a $pq \times pq$ covariance matrix $\mathbf{\Sigma}_0$ as the Kronecker product of two lower dimensional covariance matrices. When the variables are multivariate Gaussian with covariance following the KP model, the variables are said to follow a matrix normal distribution [19], [17], [20]. This model has applications in channel modeling for MIMO wireless communications [21], geostatistics [22], genomics [23], multi-task learning [24], face recognition [8], recommendation systems [9] and collaborative filtering [25]. The main difficulty in maximum likelihood estimation of structured covariances is the nonconvex optimization problem that arises. Thus, an alternating optimization approach is usually adopted. In the case where there is no missing data, an extension of the alternating optimization algorithm of Werner *et al* [18], that the authors called the flip flop (FF) algorithm, can be applied to estimate the parameters of the Kronecker product model, called KGlasso in [2].

In this paper, we assume that the covariance can be represented as a sum of Kronecker products of two lower dimensional factor matrices, where the number of terms in the summation may depend on the factor dimensions. More concretely, we assume that there are $d = pq$ variables whose covariance $\mathbf{\Sigma}_0$ has Kronecker product representation:

$$\mathbf{\Sigma}_0 = \sum_{\gamma=1}^{r} \mathbf{A}_{0,\gamma} \otimes \mathbf{B}_{0,\gamma} \tag{1}$$

where $\{\mathbf{A}_{0,\gamma}\}$ are $p \times p$ linearly independent matrices and $\{\mathbf{B}_{0,\gamma}\}$ are $q \times q$ linearly independent matrices [1]. We assume that the factor dimensions p, q are known. We note $1 \leq r \leq r_0 = \min(p^2, q^2)$ and refer to r as the *separation rank*. The model (1) is analogous to separable approximation of continuous functions [26]. It is evocative of a type of low rank principal component decomposition where the components are Kronecker products. However, the components in (1) are neither orthogonal nor normalized. The model (1) with separation rank 1 is relevant to channel modeling for MIMO wireless communications, where $\mathbf{A}_0$ is a transmit covariance matrix and $\mathbf{B}_0$ is a receive covariance matrix [21]. The model is also relevant

[1]Linear independence is understood with respect to the trace inner product defined in the space of symmetric matrices.

to other transposable models arising in recommendation systems like NetFlix and in gene expression analysis [9]. The model (1) with $r \geq 1$ also finds concrete applications in spatiotemporal MEG/EEG covariance modeling [27], [28], [29], [30] and SAR data analysis [31]. We finally note that Van Loan and Pitsianis [32] have shown that any $pq \times pq$ matrix $\mathbf{\Sigma}_0$ can be written as an orthogonal expansion of Kronecker products of the form (1), thus allowing any covariance matrix to be approximated by a bilinear decomposition of the form (1).

We propose a convex framework for estimating covariance matrices of the form (1). We call our method the Permuted Rank-penalized Least Squares (PRLS) estimator. We analyze the convergence rate of PRLS in the high dimensional setting. The main contribution of this paper is a convex optimization approach to estimating covariance matrices with KP structure of the form (1) and the derivation of tight high-dimensional MSE convergence rates as n, p and q go to infinity. For estimating separation rank r covariance matrices of the form (1), we establish that PRLS achieves high dimensional consistency with a convergence rate of $O_P\left(\frac{r(p^2+q^2+\log\max(p,q,n))}{n}\right)$. This is significantly faster than the convergence rate $O_P\left(\frac{p^2q^2}{n}\right)$ of the standard sample covariance matrix (SCM). For $r = 1$ this rate is identical to that of the FF algorithm, which fits the sample covariance matrix to a single Kronecker factor.

The PRLS method for estimating the Kronecker product expansion (1) generalizes previously studied Kronecker product covariance models [17], [19] to the case of $r > 1$. This is a fundamentally different generalization than the $r = 1$ sparse KP models proposed in [9], [2], [3], [33]. Independently in [2], [3] and [33], it was established that the high dimensional convergence rate for these sparse KP models is of order $O_P\left(\frac{(p+q)\log\max(p,q,n)}{n}\right)$. While we do not pursue the additional sparsity constraint in this paper, we speculate that sparsity can be combined with the Kronecker sum model (1), achieving improved convergence.

Advantages of the proposed PRLS covariance estimator is illustrated on both simulated and real data. The application of PRLS to the NCEP wind dataset shows that a low order Kronecker sum provides a remarkably good fit to the spatio-temporal sample covariance matrix: over 86% of all the energy is contained in the first Kronecker component of the Kronecker expansion as compared to only 41% in the principal component of the standard PCA eigen-expansion. Furthermore, by replacing the SCM in the standard linear predictor by our Kronecker sum estimator we demonstrate a 1.9 dB RMSE advantage for predicting next-day wind speeds from past measurements from the NCEP network of wind stations.

The outline of the paper is as follows. Section II introduces the notation that will be used throughout the paper. Section III introduces the PRLS covariance estimation method. Section IV presents the high-dimensional MSE convergence rate of PRLS. Section V presents numerical experiments. The technical

proofs are placed in the Appendix.

II. NOTATION

For a square matrix $\mathbf{M}$, define $|\mathbf{M}|_1 = \|\text{vec}(\mathbf{M})\|_1$ and $|\mathbf{M}|_\infty = \|\text{vec}(\mathbf{M})\|_\infty$, where $\text{vec}(\mathbf{M})$ denotes the vectorized form of $\mathbf{M}$ (concatenation of columns into a vector). $\|\mathbf{M}\|_2$ is the spectral norm of $\mathbf{M}$. $\mathbf{M}_{i,j}$ and $[\mathbf{M}]_{i,j}$ are the (i,j)th element of $\mathbf{M}$. Let the inverse transformation (from a vector to a matrix) be defined as: $\text{vec}^{-1}(\mathbf{x}) = \mathbf{X}$, where $\mathbf{x} = \text{vec}(\mathbf{X})$. Define the $pq \times pq$ permutation operator $\mathbf{K}_{p,q}$ such that $\mathbf{K}_{p,q}\text{vec}(\mathbf{N}) = \text{vec}(\mathbf{N}^T)$ for any $p \times q$ matrix $\mathbf{N}$. For a symmetric matrix $\mathbf{M}$, $\lambda(\mathbf{M})$ will denote the vector of real eigenvalues of $\mathbf{M}$ and define $\lambda_{max}(\mathbf{M}) = \|\mathbf{M}\|_2 = \max \lambda_i(\mathbf{M})$ for p.d. symmetric matrix, and $\lambda_{min}(\mathbf{M}) = \min \lambda_i(\mathbf{M})$. For any matrix $\mathbf{M}$, define the nuclear norm $\|\mathbf{M}\|_* = \sum_{l=1}^{r_M} |\sigma_l(\mathbf{M})|$, where $r_M = \text{rank}(\mathbf{M})$ and $\sigma_l(\mathbf{M})$ is the lth singular value of $\mathbf{M}$.

For a matrix $\mathbf{M}$ of size $pq \times pq$, let $\{\mathbf{M}(i,j)\}_{i,j=1}^{p}$ denote its $q \times q$ block submatrices, where each block submatrix is $\mathbf{M}(i,j) = [\mathbf{M}]_{(i-1)q+1:iq,(j-1)q+1:jq}$. Also let $\{\overline{\mathbf{M}}(k,l)\}_{k,l=1}^{q}$ denote the $p \times p$ block submatrices of the permuted matrix $\overline{\mathbf{M}} = \mathbf{K}_{p,q}^T \mathbf{M} \mathbf{K}_{p,q}$. Define the permutation operator $\mathcal{R} : \mathbb{R}^{pq \times pq} \to \mathbb{R}^{p^2 \times q^2}$ by setting the $(i-1)p+j$ row of $\mathcal{R}(\mathbf{M})$ equal to $\text{vec}(\mathbf{M}(i,j))^T$. An illustration of this permutation operator is shown in Fig. 1.

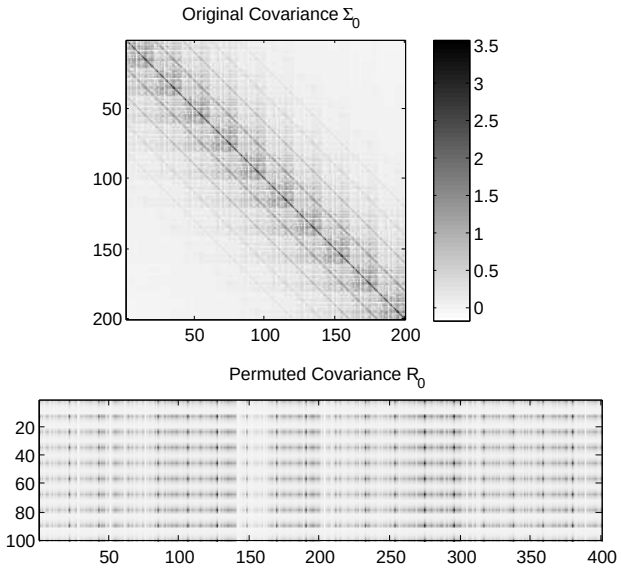

Fig. 1. Original (top) and permuted covariance (bottom) matrix. The original covariance is $\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0$, where $\mathbf{A}_0$ is a 10×10 Toeplitz matrix and $\mathbf{B}_0$ is a 20×20 unstructured p.d. matrix. Note that the permutation operator $\mathcal{R}$ maps a symmetric p.s.d. matrix $\mathbf{\Sigma}_0$ to a non-symmetric rank 1 matrix $\mathbf{R}_0 = \mathcal{R}(\mathbf{\Sigma}_0)$.

Define the set of symmetric matrices $S^p = \{\mathbf{A} \in \mathbb{R}^{p\times p} : \mathbf{A} = \mathbf{A}^T\}$, the set of symmetric positive semidefinite (psd) matrices S^p_+, and the set of symmetric positive definite (pd) matrices S^p_{++}. $\mathbf{I}_d$ is a $d \times d$ identity matrix. It can be shown that S^p_{++} is a convex set, but is not closed [34]. Note that S^p_{++} is simply the interior of the closed convex cone S^p_+.

For a subspace U, define $\mathbf{P}_U$ and $\mathbf{P}^\perp_U$ as the orthogonal projection onto U and $U^\perp$, respectively. The unit Euclidean sphere in $\mathbb{R}^{d'}$ is denoted by $\mathcal{S}^{d'-1} = \{\mathbf{x} \in \mathbb{R}^{d'} : \|\mathbf{x}\|_2 = 1\}$. Let $(x)_+ = \max(x, 0)$.

Statistical convergence rates will be denoted by the $O_P(\cdot)$ notation, which is defined as follows. Consider a sequence of real random variables $\{X_n\}_{n\in\mathbb{N}}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ and a deterministic (positive) sequence of reals $\{b_n\}_{n\in\mathbb{N}}$. By $X_n = O_P(1)$ is meant: $\sup_{n\in\mathbb{N}} \Pr(|X_n| > K) \to 0$ as $K \to \infty$, where X_n is a sequence indexed by n, for fixed p, q. The notation $X_n = O_P(b_n)$ is equivalent to $\frac{X_n}{b_n} = O_P(1)$. By $X_n = o_p(1)$ is meant $\Pr(|X_n| > \epsilon) \to 0$ as $n \to \infty$ for any $\epsilon > 0$. By $\lambda_n \asymp b_n$ is meant $c_1 \leq \frac{\lambda_n}{b_n} \leq c_2$ for all n, where $c_1, c_2 > 0$ are absolute constants.

III. Permuted Rank-penalized Least-squares

Available are n i.i.d. multivariate Gaussian observations $\{\mathbf{z}_t\}_{t=1}^n$, where $\mathbf{z}_t \in \mathbb{R}^{pq}$, having zero-mean and covariance equal to (1). A sufficient statistic for covariance estimation is the well-known sample covariance matrix (SCM):

$$\hat{\mathbf{S}}_n = \frac{1}{n}\sum_{t=1}^n \mathbf{z}_t\mathbf{z}_t^T \tag{2}$$

A penalized least-squares approach was proposed in [1] for estimating a low rank covariance matrix by solving:

$$\hat{\mathbf{\Sigma}}^\lambda_n \in \arg\min_{\mathbf{S}\in S^d_{++}} \|\hat{\mathbf{S}}_n - \mathbf{S}\|_F^2 + \lambda\|\mathbf{S}\|_* \tag{3}$$

where $\lambda > 0$ is a regularization parameter. For $\lambda = C'\|\mathbf{\Sigma}_0\|_2\sqrt{\frac{r(\mathbf{\Sigma}_0)\log(2d)}{n}}$, where $C' > 0$ is large enough, and $n \geq cr(\mathbf{\Sigma}_0)\log^2(\max(2d, n))$ for some constant $c > 0$ sufficiently large, Cor. 1 in [1] establishes a tight Frobenius norm error bound, which states that with probability $1 - \frac{1}{2p}$:

$$\|\hat{\mathbf{\Sigma}}^\lambda_n - \mathbf{\Sigma}_0\|_F^2 \leq \inf_{\mathbf{S}\succ 0} \|\mathbf{\Sigma}_0 - \mathbf{S}\|_F^2 + C\|\mathbf{\Sigma}_0\|_2^2 \text{rank}(\mathbf{S})\frac{r(\mathbf{\Sigma}_0)\log(2d)}{n}$$

where $r(\mathbf{\Sigma}_0) = \frac{\text{tr}(\mathbf{\Sigma}_0)}{\|\mathbf{\Sigma}_0\|_2} \leq \min\{\text{rank}(\mathbf{\Sigma}_0), d\}$ is the effective rank [1].

Here we propose a similar nuclear norm penalization approach to estimate low separation-rank covariance matrices. Motivated by Van Loan and Pitsianis's work [32], we propose:

$$\hat{\mathbf{R}}^\lambda_n \in \arg\min_{\mathbf{R}\in\mathbb{R}^{p^2\times q^2}} \|\hat{\mathbf{R}}_n - \mathbf{R}\|_F^2 + \lambda\|\mathbf{R}\|_* \tag{4}$$

where $\hat{\mathbf{R}}_n = \mathcal{R}(\hat{\mathbf{S}}_n)$ is the permuted SCM of size $p^2 \times q^2$. The minimum-norm problem considered in [32] is:

$$\min_{\mathbf{R}\in\mathbb{R}^{p^2\times q^2}:\text{rank}(\mathbf{R})\leq r} \|\hat{\mathbf{R}}_n - \mathbf{R}\|_F^2 \tag{5}$$

We note that (4) is a convex relaxation of (5) and is more amenable to analysis. Furthermore, we show a tradeoff between approximation error (i.e., the error induced by model mismatch between the true covariance and the model) and estimation error (i.e., the error due to finite sample size) by analyzing the solution of (4). We also note that (4) is a strictly convex problem, so there exists a unique solution that can be efficiently found using well established methods [34].

The closed form solution of (4) is given by singular value thresholding (SVT):

$$\hat{\mathbf{R}}_n^\lambda = \sum_{j=1}^{r_0} \left(\sigma_j(\hat{\mathbf{R}}_n) - \frac{\lambda}{2}\right)_+ \mathbf{u}_j\mathbf{v}_j^T \tag{6}$$

where $\mathbf{u}_j$ and $\mathbf{v}_j$ are the left and right singular vectors of $\hat{\mathbf{R}}_n$. Efficient methods of solving such problems have been recently studied in the literature [35], [36]. Although empirically observed to be fast, the computational complexity of the algorithms presented in [35] and [36] is unknown. For computation of the rank r SVD requires on the order $O(p^2q^2r)$ floating point operations (see proof of Thm. 1). However, faster probabilistic-based methods for truncated SVD take $O(p^2q^2\log(r))$ computational time [37]. Thus, the computational complexity of solving (4) scales well with respect to the designed separation rank.

The next theorem shows that the de-permuted solution is symmetric and positive definite under mild conditions.

Theorem 1. *Consider the de-permuted solution $\hat{\mathbf{\Sigma}}_n^\lambda = \mathcal{R}^{-1}(\hat{\mathbf{R}}_n^\lambda)$. The following are true:*

1) *The solution $\hat{\mathbf{\Sigma}}_n^\lambda$ is symmetric.*
2) *If $n \geq pq$, then the solution $\hat{\mathbf{\Sigma}}_n^\lambda$ is positive definite with probability 1.*

Proof: See Appendix A. ■

We believe that the PRLS estimate $\hat{\mathbf{\Sigma}}_n^\lambda$ is positive definite even if $n < pq$ for appropriately selected $\lambda > 0$. In our simulations, we always found $\hat{\mathbf{\Sigma}}_n^\lambda$ to be positive definite. We also noticed that the condition number of the PRLS estimate is order of magnitudes smaller than the SCM.

IV. High Dimensional Consistency of PRLS

In this section, we show that RPLS achieves the MSE statistical convergence rate of $O_P\left(\frac{r(p^2+q^2+\log M)}{n}\right)$. This result is clearly superior to the statistical convergence rate of the naive SCM estimator, particularly

when $p, q \to \infty$:

$$\|\hat{\mathbf{S}}_n - \mathbf{\Sigma}_0\|_F^2 = O_P\left(\frac{p^2 q^2}{n}\right). \tag{7}$$

The next result provides a deterministic relation between the spectral norm of $\hat{\mathbf{R}}_n - \mathbf{R}_0$ and the Frobenius norm of the the estimation error $\hat{\mathbf{R}}_n^{\lambda} - \mathbf{R}_0$.

Theorem 2. *Consider the convex optimization problem (4). When $\lambda \geq 2\|\hat{\mathbf{R}}_n - \mathbf{R}_0\|_2$, the following holds:*

$$\|\hat{\mathbf{R}}_n^{\lambda} - \mathbf{R}_0\|_F^2 \leq \inf_{\mathbf{R}} \left\{ \|\mathbf{R} - \mathbf{R}_0\|_F^2 + \frac{(1+\sqrt{2})^2}{4}\lambda^2 \mathit{rank}(\mathbf{R}) \right\} \tag{8}$$

Proof: See Appendix B. ■

A. High Dimensional Operator Norm Bound for Permuted Sample Covariance Matrix

In this subsection, we establish a tight bound on the spectral norm of the error matrix

$$\mathbf{\Delta}_n = \hat{\mathbf{R}}_n - \mathbf{R}_0 = \mathcal{R}(\hat{\mathbf{S}}_n - \mathbf{\Sigma}_0). \tag{9}$$

The standard strong law of large numbers implies that for fixed dimensions p, q, we have $\mathbf{\Delta}_n \to 0$ almost surely as $n \to \infty$. The next result will characterize the finite sample fluctuations of this convergence (in probability) measured by the spectral norm as a function of the sample size n and factor dimensions p, q. This result will be useful for establishing a tight bound on the Frobenius norm convergence rate of PRLS and can guide the selection of the regularization paramater in (4).

Theorem 3. *(Operator Norm Bound on Permuted SCM) Assume $\|\mathbf{\Sigma}_0\|_2 < \infty$ for all p, q and define $M = \max(p, q, n)$. Fix $\epsilon' = \frac{1}{3}$. Assume $t \geq \max(\sqrt{4C_1 \ln(1+\frac{2}{\epsilon'})}, 4C_2 \ln(1+\frac{2}{\epsilon'}))$ and $C = \max(C_1, C_2) > 0$. Then, with probability at least $1 - 2M^{-\frac{t}{4C}}$,*

$$\|\mathbf{\Delta}_n\|_2 \leq \frac{C_0 t}{1-2\epsilon'} \max\left\{ \frac{p^2+q^2+\log M}{n}, \sqrt{\frac{p^2+q^2+\log M}{n}} \right\} \tag{10}$$

for some absolute constant $C_0 > 0$ [2].

Proof: See Appendix D. ■

The proof technique is based on a large deviation inequality, derived in Lemma 2 in Appendix C, that characterizes the tail behavior of the quadratic form $\mathbf{x}^T \mathbf{\Delta}_n \mathbf{y}$ over the spheres $\mathbf{x} \in S^{p^2-1}$ and $\mathbf{y} \in S^{q^2-1}$.

[2]The constant in front of the rate can be tightened by optimizing it as a function of ϵ' over the interval $(0, 1/2)$, but is left as a finite constant here.

Using Lemma 2 and a sphere covering argument, the result of Theorem 3 follows (see Appendix D). Fig. 2 empirically validates the tightness of the bound (10) under the trivial separation rank 1 covariance $\mathbf{\Sigma}_0 = \mathbf{I}_p \otimes \mathbf{I}_q$.

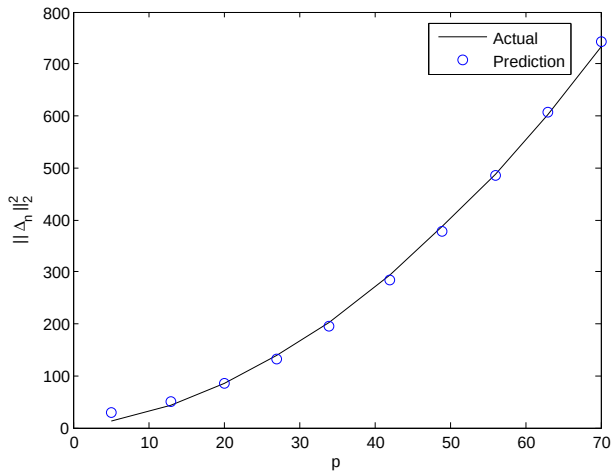

Fig. 2. Monte Carlo simulation for growth of spectral norm $\|\mathbf{\Delta}_n\|_2^2$ as a function of p for fixed $n = 10$ and $q = 5$. The predicted curve is a least-square fit of a quadratic model $y = ax^2 + b$ to the empirical curve, and is a great fit. This example shows the tightness of the probabilistic bound (10).

B. High Dimensional MSE Convergence Rate for PRLS

Using the result in Thm. 3 and the bound in Thm. 2, we next provide a tight bound on the MSE estimation error.

Theorem 4. *Define* $M = \max(p, q, n)$. *Set* $\lambda = \lambda_n = \frac{2C_0 t}{1-2\epsilon'} \max\left\{\frac{p^2+q^2+\log M}{n}, \sqrt{\frac{p^2+q^2+\log M}{n}}\right\}$ *for* $t > 0$ *large enough (see Eqn. (10)). Then, with probability at least* $1 - 2M^{-\frac{t}{4C}}$*:*

$$\|\hat{\mathbf{\Sigma}}_n^{\lambda} - \mathbf{\Sigma}_0\|_F^2 \leq \inf_{\mathbf{R}:\mathit{rank}(\mathbf{R})\leq r} \|\mathbf{R} - \mathbf{R}_0\|_F^2 + C' r \max\left\{\left(\frac{p^2+q^2+\log M}{n}\right)^2, \frac{p^2+q^2+\log M}{n}\right\} \tag{11}$$

for some absolute constant $C' > 0$.

Proof: See Appendix E. ■

When there is no model mismatch the approximation error $\inf_{\{\mathbf{R}:\text{rank}(\mathbf{R})\leq r\}} \|\mathbf{R}-\mathbf{R}_0\|_F^2$ is zero and, as a result, in the large-p, q, n asymptotic regime where $p^2+q^2+\log M = o(n)$, it follows that $\|\hat{\mathbf{\Sigma}}_n^{\lambda} - \mathbf{\Sigma}_0\|_F =$

$O_P(\sqrt{\frac{r(p^2+q^2+\log M)}{n}}) = o_p(1)$. This asymptotic MSE convergence rate of the estimated covariance to the true covariance reflects the number of degrees of freedom of the model, which is essentially of the order of $r(p^2+q^2)$ total covariance parameters. This result extends the recent high-dimensional results obtained in [2], [3] for the single Kronecker product model (i.e., $r=1$).

Moreover, we note that $r \leq r_0 = \min(p^2, q^2)$. For the case when $p \sim q$, and $r \sim r_0$, we have a fully saturated Kronecker product model and the number of model parameters are of the order $p^4 \sim d^2$, and the SCM convergence rate (7) coincides with the rate obtained in Thm. 4.

For covariance models of low separation rank-i.e., $r \ll r_0$, Thm. 4 yields that the high dimensional MSE convergence rate of PRLS can be much lower than the naive SCM convergence rate. Thus PRLS is an attractive alternative to rank-based series expansions like principal component analysis (PCA). We note that each term in the expansion $\mathbf{A}_{0,\gamma} \otimes \mathbf{B}_{0,\gamma}$ can be full-rank, while each term in the standard PCA expansion is rank 1.

Finally, we observe that Thm. 4 captures the tradeoff between estimation error and approximation error. In other words, choosing a smaller r than the true separation rank would incur a larger approximation error $\inf_{\{\mathbf{R}:\text{rank}(\mathbf{R})\leq r\}} \|\mathbf{R}-\mathbf{R}_0\|_F^2 > 0$, but smaller estimation error of the order $O_P(\frac{r(p^2+q^2+\log M)}{n})$.

C. Approximation Error

It is well known from least-squares approximation theory that the residual error can be rewritten as:

$$\inf_{\mathbf{R}:\text{rank}(\mathbf{R})\leq r} \|\mathbf{R}-\mathbf{R}_0\|_F^2 = \sum_{k=r+1}^{r_0} \sigma_k^2(\mathbf{R}_0) \tag{12}$$

Since we consider the high dimensional setting-i.e. the sample size n grows in addition to the dimensions p, q, the maximum separation rank r_0 also grows to infinity, which makes the approximation error (12) infinite. According to Theorem 4, to estimate the covariance matrix in the mean-square sense up to a bounded approximation error, we need to ensure that the sum (12) remains finite as $p, q \to \infty$. For this to occur, the singular values of $\mathbf{R}_0$ need to decay fast enough and r needs to be appropriately chosen.

We show next that the class of block-Toeplitz covariance matrices have bounded approximation error if the separation rank scaled like $\log(\max(p,q))$. To show this, we first provide a tight variational bound on the singular value spectrum of any $p^2 \times q^2$ matrix $\mathbf{R}$.

Lemma 1. *(Variational Bound on Singular Value Spectrum) Let $\mathbf{R}$ be an arbitrary matrix of size $p^2 \times q^2$. Let $\mathbf{P}_k$ be an orthogonal projection of $\mathbb{R}^{q^2}$ onto $\mathbb{R}^k$. Then, for $k = 1, \ldots, r_0 - 1$ we have:*

$$\sigma_{k+1}^2(\mathbf{R}) \leq \|(\mathbf{I}_{q^2} - \mathbf{P}_k)\mathbf{R}^T\|_2^2 \tag{13}$$

with equality iff $\mathbf{P}_k = \mathbf{V}_k\mathbf{V}_k^T$*, where* $\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ *is the singular value decomposition.*

Proof: See Appendix F. ■

Using this fundamental lemma, we can characterize the approximation error for estimating block-Toeplitz matrices with exponentially decaying off-diagonal norms. Such matrices arise, for example, as covariance matrices of multivariate stationary random processes of dimension $m$and take the block form:

$$\underbrace{\mathbf{\Sigma}_0}_{(N+1)m\times(N+1)m} = \begin{bmatrix} \mathbf{\Sigma}(0) & \mathbf{\Sigma}(1) & \dots & \mathbf{\Sigma}(N) \\ \mathbf{\Sigma}(-1) & \mathbf{\Sigma}(0) & \dots & \mathbf{\Sigma}(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Sigma}(-N) & \mathbf{\Sigma}(-N+1) & \dots & \mathbf{\Sigma}(0) \end{bmatrix} \tag{14}$$

where each submatrix is of size $m \times m$. For a zero-mean vector process $\mathbf{y} = \{\mathbf{y}(0), \dots, \mathbf{y}(N)\}$, the submatrices are given by $\mathbf{\Sigma}(\tau) = \mathbb{E}[\mathbf{y}(0)\mathbf{y}(\tau)^T]$.

Theorem 5. *Consider a block-Toeplitz p.d. matrix* $\mathbf{\Sigma}_0$ *of size* $(N+1)m \times (N+1)m$*, with* $\|\mathbf{\Sigma}(\tau)\|_F^2 \leq C'u^{2|\tau|}q$ *for all* $\tau = -N, \dots, N$ *and constant* $u \in (0,1)$*. Choose*

$$r \geq \frac{\log(pq/\epsilon)}{\log(1/u)}.$$

Then, the PRLS algorithm estimates $\mathbf{\Sigma}_0$ *up to an absolute tolerance* $\epsilon \in (0,1)$ *with convergence rate guarantee:*

$$\|\hat{\mathbf{\Sigma}}_n^\lambda - \mathbf{\Sigma}_0\|_F^2 \leq \epsilon + r\frac{p^2+q^2+\log M}{n} \tag{15}$$

for appropriately scaled λ*.*

Proof: See Appendix G. ■

The exponential norm decay condition of Thm. 5 is satisfied by a first-order vector autoregressive process:

$$\mathbf{Z}_t = \Phi\mathbf{Z}_{t-1} + \mathcal{E}_t \tag{16}$$

with $u = \|\Phi\|_2 \in (0,1)$, where $\mathbf{Z}_t$. For $\mathcal{E}_t \sim N(0, \mathbf{\Sigma}_\epsilon)$, this is a multivariate Gaussian process. Collecting data over a time horizon of size $N+1$, we concatenate these observations into a large random vector $\mathbf{z}$ of dimension $(N+1)m$, where m is the process dimension. The resulting covariance matrix has the block-Toeplitz form assumed in Thm. 5. Figure 3 shows bounds constructed using the Frobenius upper bound on the spectral norm in (13) and using the projection matrix $\mathbf{P}_k$ as discussed in the proof of Thm. 5. The bound given in the proof of Thm. 5 (in black) is shown to be linear in log-scale, thus justifying the exponential decay of the Kronecker spectrum.

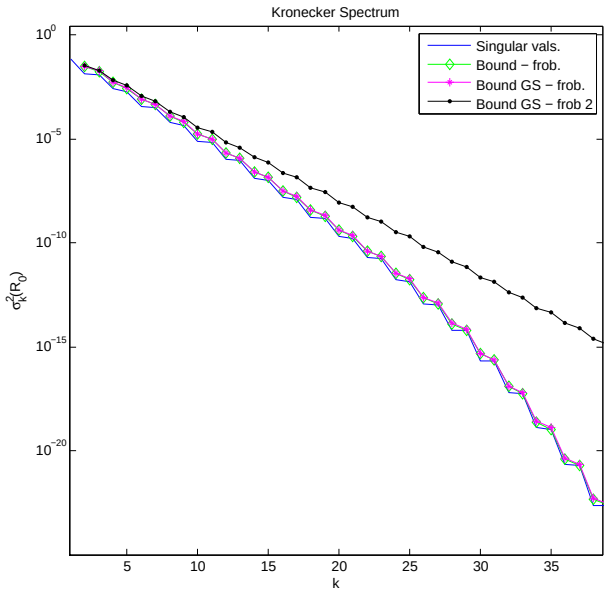

Fig. 3. Kronecker spectrum and bounds based on Lemma 1. The upper bound 'Bound - frob' (in green) is obtained using the bound (13) using the basis associated with the minimum ℓ_2 approximation error (i.e., the optimal basis computed by SVD as outlined in the equality condition of Lemma 1). The upper bound 'Bound GS - frob' (in magenta) is constructed using the variational bound (13) with projection matrix $\mathbf{P}_k$ having columns drawn from the orthonormal basis constructed in the proof of Thm. 5. The upper bound 'Bound GS - frob 2' (in black) is constructed from the bound (39) in the proof of Thm. 5.

V. Simulation Results

We consider dense positive definite matrices $\mathbf{\Sigma}_0$ of dimension $d = 625$. Taking $p = q = 25$, we note that the number of free parameters that describe each Kronecker product is of the order $p^2 + q^2 \sim p^2$, which is essentially of the same order as the number of unknown parameters required to specify each eigenvector of $\mathbf{\Sigma}_0$, i.e., $pq \sim p^2$.

A. Sum of Kronecker Product Covariance

The covariance matrix shown in Fig. 4 was constructed using (1) with $r = 3$, with each p.d. factor chosen as $\mathbf{C}\mathbf{C}^T$, where $\mathbf{C}$ is a square Gaussian random matrix. Fig. 5 shows the empirical performance of covariance matching (CM) (i.e., solution of (5) with $r = 3$), PRLS and SVT (i.e., solution of (3)). We note that the Kronecker spectrum contains only three nonzero terms while the true covariance is full rank. The PRLS spectrum is more concentrated than the eigenspectrum and, from Fig. 5, we observe PRLS outperforms covariance matching (CM), SVT and SCM across all n.

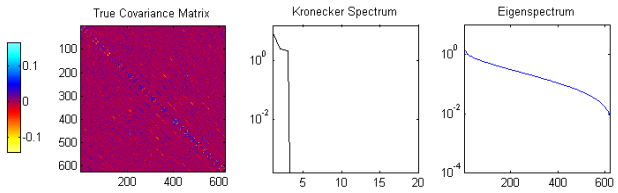

Fig. 4. Simulation A. True dense covariance is constructed using the sum of KP model (1), with $r = 3$. Left panel: True positive definite covariance matrix $\mathbf{\Sigma}_0$. Middle panel: Kronecker spectrum (eigenspectrum of $\mathbf{\Sigma}_0$ in permuted domain). Right panel: Eigenspectrum (Eigenvalues of $\mathbf{\Sigma}_0$). Note that the Kronecker spectrum is much more concentrated than the eigenspectrum.

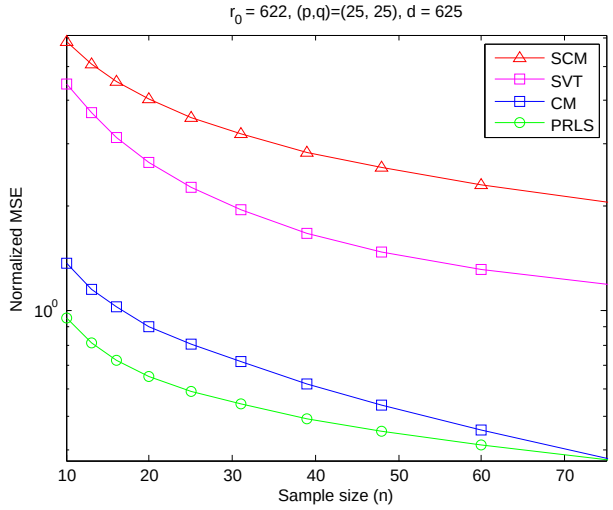

Fig. 5. Simulation A. Normalized MSE performance for true covariance matrix in Fig. 4 as a function of sample size n. PRLS outperforms CM, SVT (i.e., solution of (3)) and the standard SCM estimator. Here, $p = q = 25$ and $N_{MC} = 80$. For $n = 20$, PRLS achieves a 7.91 dB MSE reduction over SCM and SVT achieves a 1.80 dB MSE reduction over SCM.

B. Block Toeplitz Covariance

The covariance matrix shown in Fig. 6 was constructed by first generating a Gaussian random square matrix Φ of spectral norm $0.95 < 1$, and then simulating the block Toeplitz covariance for the process shown in (16). Fig. 7 compares the empirical performance of PRLS and SVT (i.e., the solution of (3) with appropriate scaling for the regularization parameter). We observe that the Kronecker product estimator performs much better than both SVT (i.e., the solution of (3)) and naive SCM estimator. This is most likely due to the fact that the repetitive block structure of Kronecker products better summarizes the covariance structure. We observe from Fig. 6 that for this block Toeplitz covariance, the Kronecker spectrum decays more rapidly (exponentially) than the eigenspectrum.

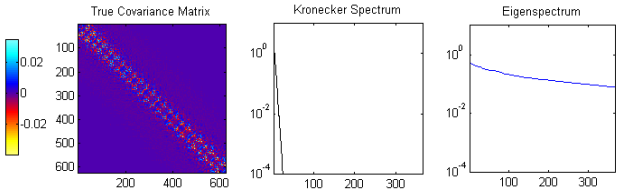

Fig. 6. Simulation B. True dense block-Toeplitz covariance matrix. Left panel: True positive definite covariance matrix $\mathbf{\Sigma}_0$. Middle panel: Kronecker spectrum (eigenspectrum of $\mathbf{\Sigma}_0$ in permuted domain). Right panel: Eigenspectrum (Eigenvalues of $\mathbf{\Sigma}_0$). Note that the Kronecker spectrum is much more concentrated than the eigenspectrum.

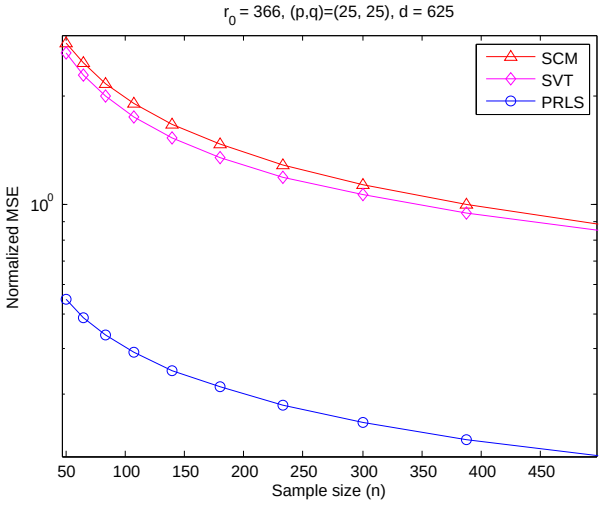

Fig. 7. Simulation B. Normalized MSE performance for covariance matrix in Fig. 6 as a function of sample size n. PRLS outperforms SVT (i.e., solution of (3)) and the standard SCM estimator. Here, $p = q = 25$ and $N_{MC} = 80$. For $n = 108$, PRLS achieves a 6.88 dB MSE reduction over SCM and SVT achieves a 0.37 dB MSE reduction over SCM. Note again that the Kronecker spectrum is much more concentrated than the eigenspectrum.

VI. Application to Wind Speed Prediction

In this section, we demonstrate the performance of PRLS in a real world application: wind speed prediction. We apply our methods to the Irish wind speed dataset and the NCEP dataset.

A. Irish Wind Speed Data

We use data consisting of time series consisting of daily average wind speed recordings during the period $1961 - 1978$ at $q = 11$ meteorological stations. This data set has many temporal coordinates, spanning a total of $n_{total} = 365 \cdot 8 = 2920$ daily average recordings of wind speed at each station. More details on this data set can be found in [38], [39], [40], [41] and it can be downloaded from Statlib

http://lib.stat.cmu.edu/datasets. We used the same square root transformation, estimated seasonal effect offset and station-specific mean offset as in [38], yielding the multiple (11) velocity measures. We used the data from years $1969-1970$ for training and the data from $1971-1978$ for testing.

The task is to predict the average velocity for the next day using the average wind velocity in each of the $p-1$ previous days. The full dimension of each observation vector is $d=pq$, and each d-dimensional observation vector is formed by concatenating the p time-consecutive q-dimensional vectors (each entry containing the velocity measure for each station) without overlapping the time segments. The SCM was estimated using data from the training period consisting of years $1969-1970$. Linear predictors over the time series were constructing by using these estimated covariance matrices in an ordinary least squares predictor. Specifically, we constructed the SCM linear predictor of all stations' wind velocity from the $p-1$ previous samples of the $q=11$ stations' time series:

$$\hat{\mathbf{v}}_t = \mathbf{\Sigma}_{2,1}\mathbf{\Sigma}_{1,1}^{-1}\mathbf{v}_{t-1:t-(p-1)} \tag{17}$$

where $\mathbf{v}_{t-1:t-(p-1)} \in \mathbb{R}^{(p-1)q}$ is the stacked wind velocities from the previous $p-1$ time instants and $\mathbf{\Sigma}_{2,1} \in \mathbb{R}^{q\times q(p-1)}$ and $\mathbf{\Sigma}_{1,1} \in \mathbb{R}^{q(p-1)\times q(p-1)}$ are submatrices of the $qp \times qp$ standard SCM:

$$\hat{\mathbf{S}}_n = \begin{bmatrix} \mathbf{\Sigma}_{1,1} & \mathbf{\Sigma}_{1,2} \\ \mathbf{\Sigma}_{2,1} & \mathbf{\Sigma}_{2,2} \end{bmatrix}$$

The PRLS predictor was similarly constructed using our proposed estimator of the $qp \times qp$ Kronecker sum covariance matrix instead of the SCM. The coefficients of each of these predictors, $\mathbf{\Sigma}_{2,1}\mathbf{\Sigma}_{1,1}^{-1}$, were subsequently applied to predict over the test set.

The predictors were tested on the data from years $1971-1978$, corresponding to $n_{test} = 365\cdot 8 = 2920$ days, as the ground truth. Using non-overlapping samples and $p=8$, we have a total of $n = \lceil \frac{365\cdot 2}{p} \rceil = 91$ training samples of full dimension $d = 88$.

Fig. 8 shows the Kronecker product factors that make up the solution of Eq. (5) with $r=1$ and the PRLS estimate. The PRLS estimate contains $r_{eff}=6$ nonzero terms in the KP expansion.

Fig. 10 shows the root mean squared error (RMSE) prediction performance over the testing period of 2920 days for the forecasts based on the standard SCM, PRLS estimator, Tyler's ML estimator [42], and regularized Tyler [42]. The PRLS estimator was implemented using a regularization parameter $\lambda_n = C\|\hat{\mathbf{S}}_n\|_2\sqrt{\frac{p^2+q^2+\log(\max(p,q,n))}{n}}$ with $C=0.13$. The constant C was chosen by optimizing the prediction RMSE on the training set over a range of regularization parameters λ parameterized by C. The regularized Tyler estimator was implemented using the data-dependent shrinkage coefficient suggested in Eqn. (13)

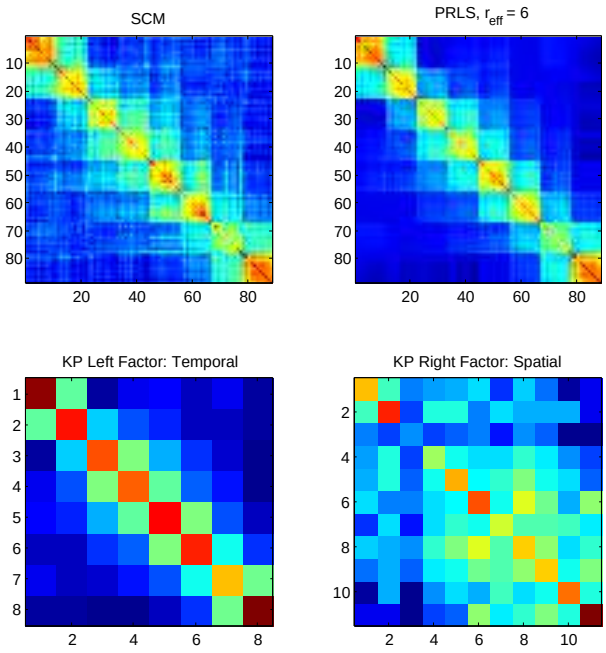

Fig. 8. Irish wind speed data: Sample covariance matrix (SCM) (top left), PRLS covariance estimate (top right), temporal Kronecker factor for first KP component (bottom left) and spatial Kronecker factor for first KP component (bottom right).

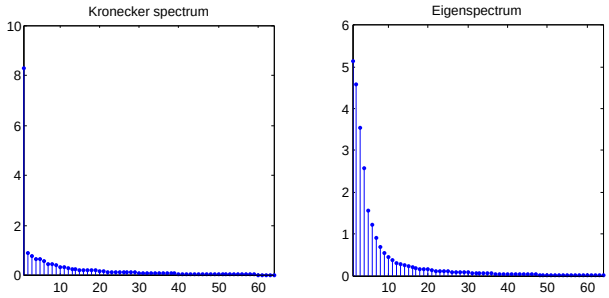

Fig. 9. Irish wind speed data: Kronecker spectrum of SCM (left) and Eigenspectrum of SCM (right). The first and second KP components contain 94.60% and 1.07% of the spectrum energy. The first and second eigenvectors contain 36.28% and 28.76% of the spectrum energy. The KP spectrum is more compact than the eigenspectrum.

in [42]. Fig. 11 shows a sample period of 150 days. We observe that PRLS tracks the actual wind speed better than the SCM-based predictor does.

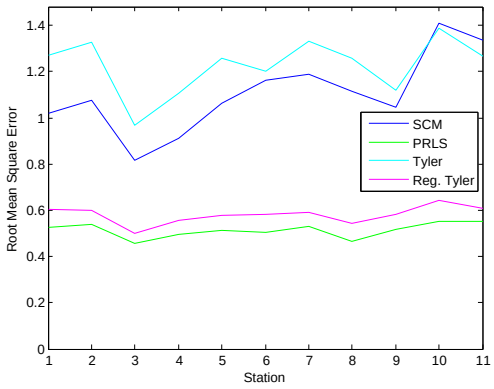

Fig. 10. Irish wind speed data: RMSE prediction performance across q stations for linear estimators using SCM (blue), PRLS (green), Tyler's MLE (cyan) and regularized Tyler (magenta). PRLS achieves an average reduction in RMSE of 3.32 dB as compared to SCM (averaged across stations) and regularized Tyler achieves an average RMSE reduction of 2.79 dB as compared to SCM.

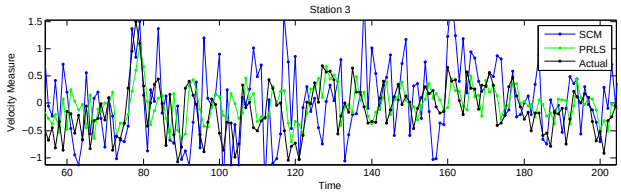

Fig. 11. Irish wind speed data: Prediction performance for linear estimators using SCM (blue) and PRLS (green) for a time interval of 150 days. The actual (ground truth) wind speeds are shown in black. PRLS offers better tracking performance as compared to SCM.

B. NCEP Wind Speed Data

We use data representative of the wind conditions in the lower troposphere (surface data at .995 sigma level) for the global grid (90°N - 90°S, 0°E - 357.5°E). We obtained the data from the National Centers for Environmental Prediction reanalysis project (Kalnay et al. [43]), available at the NOAA website ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis.dailyavgs/surface. Daily averages of U (east-west) and V (north-south) wind components were collected using a station grid of size 144×73 (2.5 degree latitude $\times$ 2.5 degree longitude global grid) over the years $1948 - 2012$. The wind speed is computed by taking the magnitude of the wind vector.

1) Continental US Region: We considered a 10×10 grid of stations, corresponding to latitude range 25°N-47.5°N and longitude range 125°W-97.5°W. For this selection of variables, $q = 10 \cdot 10 = 100$ is the total number of stations and $p - 1 = 7$ is the prediction time lag. We preprocessed the raw data using

the detrending procedure outlined in Haslett et al. [38]. More specifically, we first performed a square root transformation, then estimated and subtracted the station-specific means from the data and finally estimated and subtracted the seasonal effect (see Fig. 12). The resulting features/observations are called the velocity measures [38]. The SCM was estimated using data from the training period consisting of

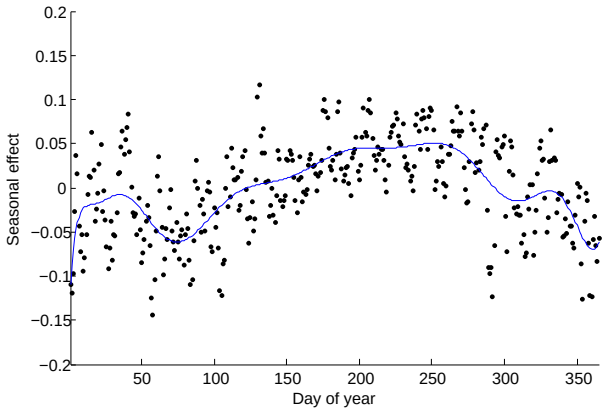

Fig. 12. NCEP wind speed data (Continental US): Seasonal effect as a function of day of the year. A 14th order polynomial is fit by the least squares method to the average of the square root of the daily mean wind speeds over all stations and over all training years.

years $2003-2007$. Since the SCM is not full rank, the linear preictor (17) was implemented with the Moore-Penrose pseudo-inverse of $\mathbf{\Sigma}_{1,1}$. The predictors were tested on the data from years $2008-2012$ as the ground truth. Using non-overlapping samples and $p=8$, we have a total of $n=\lceil\frac{365\cdot 5}{p}\rceil=228$ training samples of full dimension $d=800$.

Fig. 13 shows the Kronecker product factors that make up the solution of Eq. (5) with $r=1$ and the PRLS covariance estimate. The PRLS estimate contains $r_{eff}=6$ nonzero terms in the KP expansion.

Fig. 15 shows the root mean squared error (RMSE) prediction performance over the testing period of 1825 days for the forecasts based on the standard SCM, PRLS, and regularized Tyler [42]. The PRLS estimator was implemented using a regularization parameter $\lambda_n = C\|\hat{\mathbf{S}}_n\|_2\sqrt{\frac{p^2+q^2+\log(\max(p,q,n))}{n}}$ with $C=0.036$. The constant C was chosen by optimizing the prediction RMSE on the training set over a range of regularization parameters λ parameterized by C (as in Irish wind speed data set). Fig. 16 shows a sample period of 150 days. It is observed that SCM has unstable performance, while the Kronecker product estimator offers better tracking of the wind speeds.

2) Arctic Ocean Region: We considered a 10×10 grid of stations, corresponding to latitude range $90°$N-$67.5°$N and longitude range $0°$E-$22.5°$E. For this selection of variables, $q=10\cdot 10=100$ is the

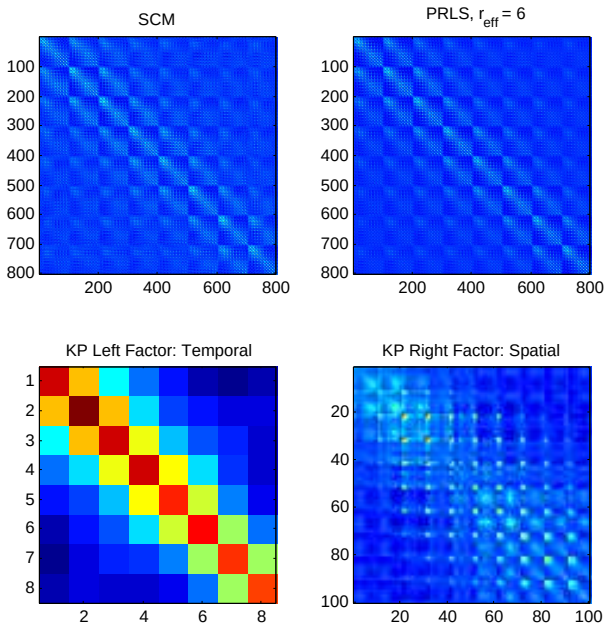

Fig. 13. NCEP wind speed data (Continental US): Sample covariance matrix (SCM) (top left), PRLS covariance estimate (top right), temporal Kronecker factor for first KP component (bottom left) and spatial Kronecker factor for first KP component (bottom right).

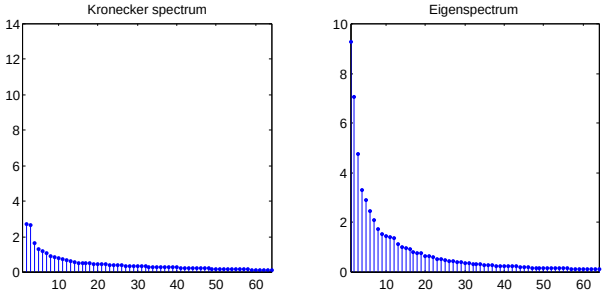

Fig. 14. NCEP wind speed data (Continental US): Kronecker spectrum of SCM (left) and Eigenspectrum of SCM (right). The first and second KP components contain 85.88% and 3.48% of the spectrum energy. The first and second eigenvectors contain 40.93% and 23.82% of the spectrum energy. The KP spectrum is more compact than the eigenspectrum.

total number of stations and $p - 1 = 7$ is the prediction time lag. We preprocessed the raw data using the detrending procedure outlined in Haslett et al. [38]. More specifically, we first performed a square root transformation, then estimated and subtracted the station-specific means from the data and finally estimated and subtracted the seasonal effect (see Fig. 17). The resulting features/observations are called

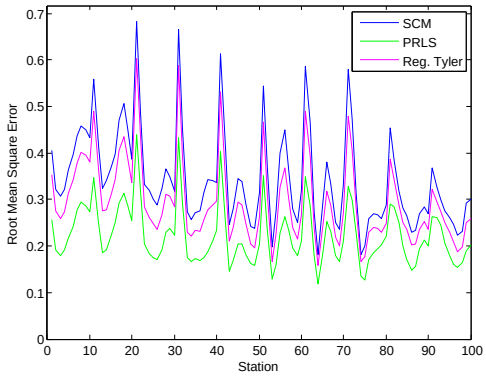

Fig. 15. NCEP wind speed data (Continental US): RMSE prediction performance across q stations for linear estimators using SCM (blue) and PRLS (green). PRLS achieves an average reduction in RMSE of 1.90 dB as compared to SCM (averaged across stations) and regularized Tyler achieves an average RMSE reduction of 0.66 dB as compared to SCM.

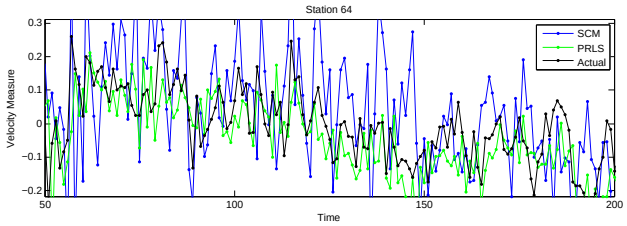

Fig. 16. NCEP wind speed data (Continental US): Prediction performance for linear estimators using SCM (blue) and PRLS (green) for a time interval of 150 days. The actual (ground truth) wind speeds are shown in black. PRLS offers better tracking performance as compared to SCM.

the velocity measures [38]. The SCM was estimated using data from the training period consisting of years $2003-2007$. Since the SCM is not full rank, the linear preictor (17) was implemented with the Moore-Penrose pseudo-inverse of $\mathbf{\Sigma}_{1,1}$. The predictors were tested on the data from years $2008-2012$ as the ground truth. Using non-overlapping samples and $p=8$, we have a total of $n=\lceil\frac{365\cdot 5}{p}\rceil=228$ training samples of full dimension $d=800$.

Fig. 18 shows the Kronecker product factors that make up the solution of Eq. (5) with $r=1$ and the PRLS covariance estimate. The PRLS estimate contains $r_{eff}=2$ nonzero terms in the KP expansion.

Fig. 20 shows the root mean squared error (RMSE) prediction performance over the testing period of 1825 days for the forecasts based on the standard SCM, PRLS, and regularized Tyler [42]. The PRLS estimator was implemented using a regularization parameter $\lambda_n = C\|\hat{\mathbf{S}}_n\|_2\sqrt{\frac{p^2+q^2+\log(\max(p,q,n))}{n}}$ with $C=0.073$. The constant C was chosen by optimizing the prediction RMSE on the training set over a

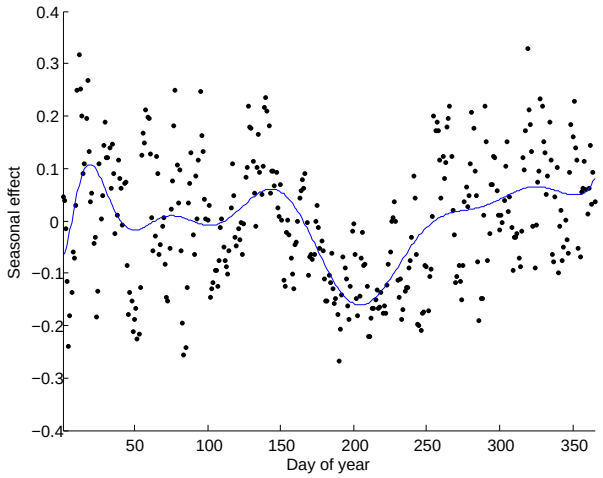

Fig. 17. NCEP wind speed data (Arctic Ocean): Seasonal effect as a function of day of the year. A 14th order polynomial is fit by the least squares method to the average of the square root of the daily mean wind speeds over all stations and over all training years.

range of regularization parameters λ parameterized by C (as in Irish wind speed data set). Fig. 21 shows a sample period of 150 days. It is observed that SCM has unstable performance, while the Kronecker product estimator offers better tracking of the wind speeds.

VII. Conclusion

We have introduced a framework for covariance estimation based on separation rank decompositions using a series of Kronecker product factors. We proposed a least-squares estimator in a permuted linear space with nuclear norm penalization, named PRLS. We established high dimensional consistency for PRLS with guaranteed rates of convergence. The analysis shows that for low separation rank covariance models, our proposed method outperforms the standard SCM estimator. For the class of block-Toeplitz matrices with exponentially decaying off-diagonal norms, we showed that the separation rank is small, and specialized our convergence bounds to this class. We also presented synthetic simulations that showed the benefits of our methods.

We emphasize that our analysis is in general non-asymptotic, in the sense that probabilistic bounds are derived that hold with a certain probability and this probability becomes higher as the number of sample and/or variables tend to infinity.

As a real world application we demonstrated the performance of the proposed Kronecker product-based estimator in wind speed prediction using an Irish wind speed dataset and a recent US NCEP dataset.

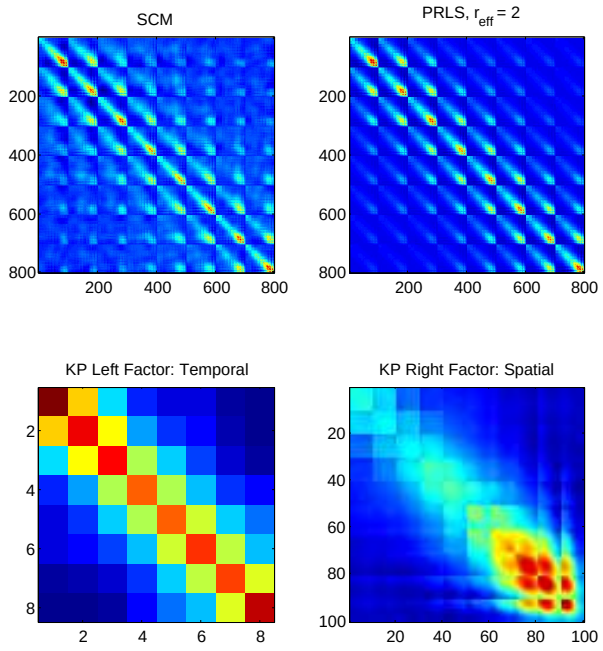

Fig. 18. NCEP wind speed data (Arctic Ocean): Sample covariance matrix (SCM) (top left), PRLS covariance estimate (top right), temporal Kronecker factor for first KP component (bottom left) and spatial Kronecker factor for first KP component (bottom right).

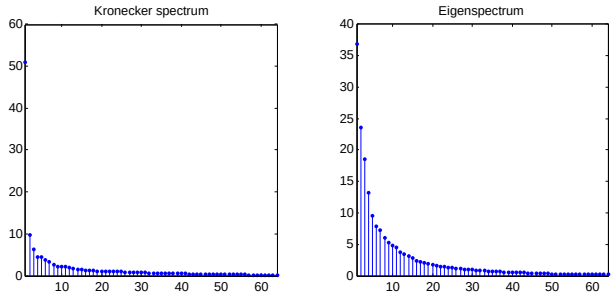

Fig. 19. NCEP wind speed data (Arctic Ocean): Kronecker spectrum of SCM (left) and Eigenspectrum of SCM (right). The first and second KP components contain 91.12% and 3.28% of the spectrum energy. The first and second eigenvectors contain 47.99% and 19.68% of the spectrum energy. The KP spectrum is more compact than the eigenspectrum.

Implementation of a standard covariance-based prediction scheme using our Kronecker product estimator achieved performance gains as compared to standard with respect to previously proposed covariance-based predictors.

There are several questions that remain open and are worthy of additional study. First, while the

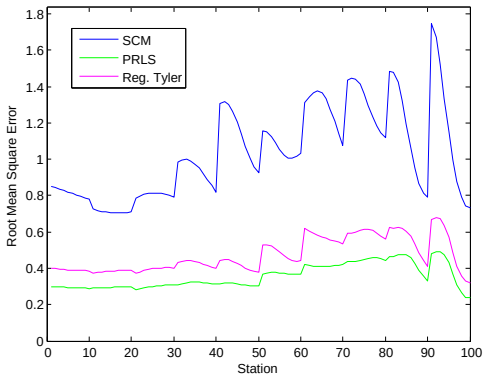

Fig. 20. NCEP wind speed data (Arctic Ocean): RMSE prediction performance across q stations for linear estimators using SCM (blue) and PRLS (green). PRLS achieves an average reduction in RMSE of 4.64 dB as compared to SCM (averaged across stations) and regularized Tyler achieves an average RMSE reduction of 3.41 dB as compared to SCM.

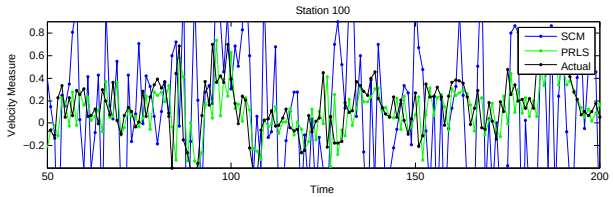

Fig. 21. NCEP wind speed data (Arctic Ocean): Prediction performance for linear estimators using SCM (blue) and PRLS (green) for a time interval of 150 days. The actual (ground truth) wind speeds are shown in black. PRLS offers better tracking performance as compared to SCM.

proposed penalized least squares Kronecker sum approximation yields a unique solution, the solution requires specification of the parameter λ, which specifies both the separation rank, and the amount of spectral shrinkage in the approximation. It would be worthwhile to investigate optimal or consistent methods of choosing this regularization parameter, e.g. using Stein's theory of unbiased risk minimization. Second, while we have proven positive definiteness of the Kronecker sum approximation when the number of samples is greater than than the variable dimension in our experiments we have observed that positive definiteness is preserved more generally. The positive definiteness conditions should be investigated further. Finally, maximum likelihood estimation of Kronecker sum covariance and inverse covariance matrices is a worthwhile open problem.

ACKNOWLEDGEMENT

The research reported in this paper was supported in part by ARO grant W911NF-11-1-0391.

APPENDIX A
PROOF OF THEOREM 1

Proof:

1) Note that the singular value decomposition of $\hat{\mathbf{R}}_n = \mathcal{R}(\hat{\mathbf{S}}_n)$ is equivalent to the solution of the minimum norm problem:

$$\min_{\{\mathbf{A}_k, \mathbf{B}_k\}_k} \|\hat{\mathbf{S}}_n - \sum_{k=1}^{r} \mathbf{A}_k \otimes \mathbf{B}_k\|_F^2 \tag{18}$$

This problem is solved by a permuted SVD [32], which can be equivalently formulated as a successive rank-1 approximation problem in a different linear space and each rank-1 approximation problem can be written as an iterative system [29]. After some algebra, this sequence of iterative systems takes the form:

$$\begin{aligned}
\mathbf{A}_1 &= \hat{\mathbf{A}}(\mathbf{B}_1) \\
\mathbf{B}_1 &= \hat{\mathbf{B}}(\mathbf{A}_1) \\
\mathbf{A}_u &= \hat{\mathbf{A}}(\mathbf{B}_u) - \sum_{\alpha=1}^{u-1} < \mathbf{A}_\alpha, \hat{\mathbf{A}}(\mathbf{B}_u) > \mathbf{A}_\alpha, 2 \leq u \leq r \\
\mathbf{B}_u &= \hat{\mathbf{B}}(\mathbf{A}_u) - \sum_{\beta=1}^{u-1} \frac{< \mathbf{B}_\beta, \hat{\mathbf{B}}(\mathbf{A}_u) >}{\|\mathbf{B}_\beta\|_F^2} \mathbf{B}_\beta, 2 \leq u \leq r
\end{aligned} \tag{19}$$

for any initial p.d. starting matrix $\mathbf{B}_u, u \geq 1$. The operators $\hat{\mathbf{A}}(\cdot)$ and $\hat{\mathbf{B}}(\cdot)$ are defined as:

$$\hat{\mathbf{A}}(\mathbf{B}) = \frac{1}{\|\mathbf{B}\|_F^2} \sum_{k,l=1}^{q} [\mathbf{B}]_{k,l} \bar{\hat{\mathbf{S}}}_n(l,k) \tag{20}$$

$$\hat{\mathbf{B}}(\mathbf{A}) = \frac{1}{\|\mathbf{A}\|_F^2} \sum_{i,j=1}^{p} [\mathbf{A}]_{i,j} \hat{\mathbf{S}}_n(j,i) \tag{21}$$

Since the operators $\hat{\mathbf{A}}(\cdot)$ and $\hat{\mathbf{B}}(\cdot)$ inherit symmetry from $\hat{\mathbf{S}}_n$, the proof is complete.

2) Recall that the SCM $\hat{\mathbf{S}}_n$ is positive definite with probability 1 if $n \geq pq$. First, consider the minimum norm problem (18). From the first part of the proposition, it follows that the factors $\mathbf{A}_k$ and $\mathbf{B}_k$ must be symmetric. If we show that there exists a solution to (18) with p.d. Kronecker factors, then the weighted sum with positive scalars is also p.d. and as a result, the PRLS solution given by $\hat{\mathbf{\Sigma}}_n^\lambda = \sum_{k=1}^{r_0} \left(\sigma_k(\hat{\mathbf{R}}_n) - \frac{\lambda}{2}\right)_+ \mathbf{U}_k \otimes \mathbf{V}_k$ is positive definite (see Eqn. 6).
Fix $l \in \{1, \ldots, r\}$. We will show that $\mathbf{A}_l$ and $\mathbf{B}_l$ are p.d. matrices. From eigendecomposition, it follows that there exist orthonormal matrices $\mathbf{\Psi}_l$ and $\mathbf{\Xi}_l$ and diagonal matrices $\mathbf{D}_l = \text{diag}(d_l^1, \ldots, d_l^p)$

and $\mathbf{\Lambda}_l = \text{diag}(\lambda_l^1, \ldots, \lambda_l^q)$ such that:

$$\mathbf{D}_l = \mathbf{\Psi}_l^T \mathbf{A}_l \mathbf{\Psi}_l$$

$$\mathbf{\Lambda}_l = \mathbf{\Xi}_l^T \mathbf{B}_l \mathbf{\Xi}_l$$

Set $\mathbf{Q}_l = \mathbf{\Psi}_l \otimes \mathbf{\Xi}_l$. Define $\mathbf{F}_l = \mathbf{Q}_l^T \hat{\mathbf{S}}_n \mathbf{Q}_l$. The objective (18) can be rewritten as:

$$\begin{aligned}
&\|\hat{\mathbf{S}}_n - \sum_{k=1}^{r} \mathbf{A}_k \otimes \mathbf{B}_k\|_F^2 \\
&= \|\mathbf{Q}_l^T \left(\hat{\mathbf{S}}_n - \sum_{k=1}^{r} \mathbf{A}_k \otimes \mathbf{B}_k \right) \mathbf{Q}_l\|_F^2 \\
&= \|\mathbf{F}_l - \sum_{k=1}^{r} \mathbf{Q}_l^T (\mathbf{A}_k \otimes \mathbf{B}_k) \mathbf{Q}_l\|_F^2 \\
&= \| \underbrace{\mathbf{F}_l - \sum_{k \neq l} (\mathbf{\Psi}_l^T \mathbf{A}_k \mathbf{\Psi}_l) \otimes (\mathbf{\Xi}_l^T \mathbf{B}_k \mathbf{\Xi}_l)}_{\mathbf{M}_l} \\
&\quad - (\mathbf{\Psi}_l^T \mathbf{A}_l \mathbf{\Psi}_l) \otimes (\mathbf{\Xi}_l^T \mathbf{B}_l \mathbf{\Xi}_l)\|_F^2 \\
&= \|\mathbf{M}_l - \mathbf{D}_l \otimes \mathbf{\Lambda}_l\|_F^2 \\
&= \|\mathbf{M}_l\|_F^2 + \|\mathbf{D}_l \otimes \mathbf{\Lambda}_l\|_F^2 - 2\text{tr}\left(\mathbf{F}_l (\mathbf{D}_l \otimes \mathbf{\Lambda}_l)\right) \\
&\quad + 2 \sum_{k \neq l} \text{tr}((\mathbf{\Psi}_l^T \mathbf{A}_k \mathbf{\Psi}_l \otimes \mathbf{\Xi}_l^T \mathbf{B}_k \mathbf{\Xi}_l)(\mathbf{D}_l \otimes \mathbf{\Lambda}_l)) \\
&= \|\mathbf{M}_l\|_F^2 - \|\mathbf{F}_l\|_F^2 + \|\mathbf{F}_l - \mathbf{D}_l \otimes \mathbf{\Lambda}_l\|_F^2 \\
&\quad + 2 \sum_{k \neq l} \text{tr}(\mathbf{B}_k \mathbf{B}_l) \text{tr}(\mathbf{C}_k \mathbf{C}_l) \\
&= \|\mathbf{M}_l\|_F^2 - \|\mathbf{F}_l\|_F^2 + \|\mathbf{F}_l - \text{diag}(\mathbf{F}_l)\|_F^2 \\
&\quad + \|\text{diag}(\mathbf{F}_l) - \mathbf{D}_l \otimes \mathbf{\Lambda}_l\|_F^2
\end{aligned}$$

where we used the orthogonality of Kronecker factors induced by the SVD solution in the last step. We note that the term $\|\mathbf{M}_l\|_F^2 - \|\mathbf{F}_l\|_F^2 + \|\mathbf{F}_l - \text{diag}(\mathbf{F}_l)\|_F^2$ is independent of $\mathbf{D}_l, \mathbf{\Lambda}_l$. The positive definiteness of $\mathbf{S}_n$ implies that the diagonal elements of $\mathbf{F}_l$ are all positive. Let $\text{diag}(\mathbf{F}_l) =$

$\text{diag}(\{f_{(i-1)q+j}\}_{i,j}) > 0$. Direct algebra yields:

$$\begin{aligned}
&\|\text{diag}(\mathbf{F}_l) - \mathbf{D}_l \otimes \mathbf{\Lambda}_l\|_F^2 \\
&\quad = \sum_{i=1}^{p}\sum_{j=1}^{q}(f_{(i-1)q+j} - d_i\lambda_j)^2 \\
&\quad = \sum_{i=1}^{p}\sum_{j=1}^{q}(f_{(i-1)q+j} - |d_i||\lambda_j|)^2 \\
&\qquad + 2\sum_{i=1}^{p}\sum_{j=1}^{q} f_{(i-1)q+j}(|d_i||\lambda_j| - d_i\lambda_j)
\end{aligned}$$

We note that the term $\sum_{i=1}^{p}\sum_{j=1}^{q}(f_{(i-1)q+j} - |d_i||\lambda_j|)^2$ is invariant to any sign changes of the eigenvalues $\{d_i, \lambda_j\}_{i,j}$. By contradiction, it follows that the eigenvalues $\{d_i\}$ and $\{\lambda_j\}$ must all have the same sign (if not, then the minimum norm is not achieved by $(\mathbf{A}_l, \mathbf{B}_l)$). Without loss of generality (since $\mathbf{A}_l \otimes \mathbf{B}_l = (-\mathbf{A}_l) \otimes (-\mathbf{B}_l)$, we assume the sign is positive. We conclude that there exist p.d. matrices $(\mathbf{A}_l, \mathbf{B}_l)$ that achieve the minimum norm. Unfixing l and using the convexity of the PRLS objective (4) concludes the proof.

■

APPENDIX B
PROOF OF THEOREM 2

Proof: The proof generalizes Thm. 1 in [1] to nonsquare matrices. A necessary and sufficient condition for the minimizer of (4) is that there exists a $\hat{\mathbf{V}} \in \partial\|\hat{\mathbf{R}}^{\lambda}\|_*$ such that:

$$< 2(\hat{\mathbf{R}}^{\lambda} - \hat{\mathbf{R}}_n) + \lambda\hat{\mathbf{V}}, \hat{\mathbf{R}}^{\lambda} - \mathbf{R} >\leq 0 \tag{22}$$

for all $\mathbf{R}$. From (22), we obtain for any $\mathbf{V} \in \partial\|\mathbf{R}\|_1$:

$$\begin{aligned}
&2 < \hat{\mathbf{R}}^{\lambda} - \mathbf{R}_0, \hat{\mathbf{R}}^{\lambda} - \mathbf{R} > + \lambda < \hat{\mathbf{V}} - \mathbf{V}, \hat{\mathbf{R}}^{\lambda} - \mathbf{R} > \\
&\leq -\lambda < \mathbf{V}, \hat{\mathbf{R}}^{\lambda} - \mathbf{R} > + 2 < \hat{\mathbf{R}}_n - \mathbf{R}_0, \hat{\mathbf{R}}^{\lambda} - \mathbf{R} >
\end{aligned} \tag{23}$$

The monotonicity of subdifferentials of convex functions implies:

$$< \hat{\mathbf{V}} - \mathbf{V}, \hat{\mathbf{R}}^{\lambda} - \mathbf{R} >\geq 0 \tag{24}$$

From Example 2 in [44], we have the characterization of the subdifferential of a nuclear norm of a nonsquare matrix:

$$\partial\|\mathbf{R}\|_* = \left\{ \sum_{j=1}^{r} \mathbf{u}_j(\mathbf{R})\mathbf{v}_j(\mathbf{R})^T + \mathbf{P}_U^{\perp}\mathbf{W}\mathbf{P}_V^{\perp} : \|\mathbf{W}\|_2 \leq 1 \right\}$$

where $r = \text{rank}(\mathbf{R})$, $U = \text{span}\{\mathbf{u}_j\}$ and $V = \text{span}\{\mathbf{v}_j\}$. Thus, for $\mathbf{R} = \sum_{j=1}^{r} \sigma_j(\mathbf{R})\mathbf{u}_j\mathbf{v}_j^T$, $r = \text{rank}(\mathbf{R})$, we can write:

$$\mathbf{V} = \sum_{j=1}^{r} \mathbf{u}_j\mathbf{v}_j^T + \mathbf{P}_U^{\perp}\mathbf{W}\mathbf{P}_V^{\perp} \tag{25}$$

where $\mathbf{W}$ can be chosen such that $\|\mathbf{W}\|_2 \leq 1$ and

$$< \mathbf{P}_U^{\perp}\mathbf{W}\mathbf{P}_V^{\perp}, \hat{\mathbf{R}}^{\lambda} - \mathbf{R} >= \|\mathbf{P}_U^{\perp}\hat{\mathbf{R}}^{\lambda}\mathbf{P}_V^{\perp}\|_* \tag{26}$$

Next, note the equality:

$$\begin{aligned} &\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}_0\|_F^2 + \|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\|_F^2 - \|\mathbf{R} - \mathbf{R}_0\|_F^2 \\ &\qquad = 2 < \hat{\mathbf{R}}^{\lambda} - \mathbf{R}_0, \hat{\mathbf{R}}^{\lambda} - \mathbf{R} > \end{aligned} \tag{27}$$

Using (24), (26) and (27) in (23), we obtain:

$$\begin{aligned} &\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}_0\|_F^2 + \|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\|_F^2 + \lambda\|\mathbf{P}_U^{\perp}\hat{\mathbf{R}}^{\lambda}\mathbf{P}_V^{\perp}\|_* \\ &\leq \|\mathbf{R} - \mathbf{R}_0\|_F^2 + \lambda < \sum_{j=1}^{r} \mathbf{u}_j\mathbf{v}_j^T, -(\hat{\mathbf{R}}^{\lambda} - \mathbf{R}) > \\ &\quad + 2 < \hat{\mathbf{R}}_n - \mathbf{R}_0, \hat{\mathbf{R}}^{\lambda} - \mathbf{R} > \end{aligned} \tag{28}$$

From trace duality, we have:

$$\begin{aligned} &< \sum_{j=1}^{r} \mathbf{u}_j\mathbf{v}_j^T, -(\hat{\mathbf{R}}^{\lambda} - \mathbf{R}) > \\ &\quad =< \mathbf{P}_U \sum_{j=1}^{r} \mathbf{u}_j\mathbf{v}_j^T\mathbf{P}_V, -(\hat{\mathbf{R}}^{\lambda} - \mathbf{R}) > \\ &\quad \leq \|\sum_{j=1}^{r} \mathbf{u}_j\mathbf{v}_j^T\|_2\|\mathbf{P}_U^T(\hat{\mathbf{R}}^{\lambda} - \mathbf{R})\mathbf{P}_V^T\|_* \\ &\quad = \|\mathbf{P}_U(\hat{\mathbf{R}}^{\lambda} - \mathbf{R})\mathbf{P}_V\|_* \end{aligned}$$

where we used the symmetry of projection matrices. Using this bound in (28), we obtain:

$$\begin{aligned} &\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}_0\|_F^2 + \|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\|_F^2 + \lambda\|\mathbf{P}_U^{\perp}\hat{\mathbf{R}}^{\lambda}\mathbf{P}_V^{\perp}\|_* \\ &\leq \|\mathbf{R} - \mathbf{R}_0\|_F^2 + \lambda\|\mathbf{P}_U(\hat{\mathbf{R}}^{\lambda} - \mathbf{R})\mathbf{P}_V\|_* \\ &\quad + 2 < \mathbf{\Delta}_n, \hat{\mathbf{R}}^{\lambda} - \mathbf{R} > \end{aligned} \tag{29}$$

where $\boldsymbol{\Delta}_n = \hat{\mathbf{R}}_n - \mathbf{R}_0$. Define the orthogonal projection of $\mathbf{R}$ onto the outer product span of U and V as $\mathcal{P}_{U,V}(\mathbf{R}) = \mathbf{R} - \mathbf{P}_U^{\perp}\mathbf{R}\mathbf{P}_V^{\perp}$. Then, we decompose:

$$\begin{aligned} < \boldsymbol{\Delta}_n, \hat{\mathbf{R}}^{\lambda} - \mathbf{R} > =& < \boldsymbol{\Delta}_n, \mathcal{P}_{U,V}\left(\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\right) > \\ &+ < \boldsymbol{\Delta}_n, \mathbf{P}_U^{\perp}(\hat{\mathbf{R}}^{\lambda} - \mathbf{R})\mathbf{P}_V^{\perp} > \end{aligned}$$

By the Cauchy-Schwarz inequality and trace-duality:

$$\begin{aligned} \|\mathbf{P}_U(\hat{\mathbf{R}}^{\lambda} - \mathbf{R})\mathbf{P}_V\|_* &\leq \sqrt{\text{rank}(\mathbf{R})}\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\|_F \\ | < \boldsymbol{\Delta}_n, \mathcal{P}_{U,V}(\hat{\mathbf{R}}^{\lambda} - \mathbf{R}) > | &\leq \|\boldsymbol{\Delta}_n\|_2 \|\mathcal{P}_{U,V}(\hat{\mathbf{R}}^{\lambda} - \mathbf{R})\|_* \\ &\leq \|\boldsymbol{\Delta}_n\|_2 \sqrt{2\text{rank}(\mathbf{R})}\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\|_F \\ | < \boldsymbol{\Delta}_n, \mathbf{P}_U^{\perp}(\hat{\mathbf{R}}^{\lambda} - \mathbf{R})\mathbf{P}_V^{\perp} > | &\leq \|\boldsymbol{\Delta}_n\|_2 \|\mathbf{P}_U^{\perp}\hat{\mathbf{R}}^{\lambda}\mathbf{P}_V^{\perp}\|_* \end{aligned}$$

where we used $\mathbf{P}_U^{\perp}\mathbf{R}\mathbf{P}_V^{\perp} = 0$. Using these bounds in (29), we further obtain:

$$\begin{aligned} &\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}_0\|_F^2 + \|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\|_F^2 + (\lambda - 2\|\boldsymbol{\Delta}_n\|_2)\|\mathbf{P}_U^{\perp}\hat{\mathbf{R}}^{\lambda}\mathbf{P}_V^{\perp}\|_* \\ &\leq \|\mathbf{R} - \mathbf{R}_0\|_F^2 + ((2\sqrt{2}\|\boldsymbol{\Delta}_n\|_2 + \lambda)\sqrt{r})(\sqrt{\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}\|_F^2}) \end{aligned} \tag{30}$$

Using the arithmetic-mean geometric-mean inequality in the RHS of (30) and the assumption $\lambda \geq 2\|\boldsymbol{\Delta}_n\|_2$, we obtain:

$$\|\hat{\mathbf{R}}^{\lambda} - \mathbf{R}_0\|_F^2 \leq \|\mathbf{R} - \mathbf{R}_0\|_F^2 + \frac{\lambda^2(1+\sqrt{2})^2}{4}r$$

This concludes the proof. ■

APPENDIX C
LEMMA 2

Lemma 2. *(Concentration of Measure for Coupled Gaussian Chaos) Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be arbitrary unit-Frobenius norm matrices and let* $\mathbf{x} \in \mathbb{R}^{p^2}$ *and* $\mathbf{y} \in \mathbb{R}^{q^2}$ *be reshaped versions of* $\mathbf{X}$ *and* $\mathbf{Y}$*. In the SCM (2) assume that* $\{\mathbf{z}_t\}$ *are i.i.d. multivariate normal* $\mathbf{z}_t \sim N(0, \boldsymbol{\Sigma}_0)$*. Recall* $\boldsymbol{\Delta}_n$ *in (9). For all* $\tau \geq 0$*:*

$$\mathbb{P}(|\mathbf{x}^T\boldsymbol{\Delta}_n\mathbf{y}| \geq \tau) \leq 2\exp\left(\frac{-n\tau^2/2}{C_1\|\boldsymbol{\Sigma}_0\|_2^2 + C_2\|\boldsymbol{\Sigma}_0\|_2\tau}\right) \tag{31}$$

where $C_1 = \frac{4e}{\sqrt{6\pi}} \approx 2.5044$ *and* $C_2 = e\sqrt{2} \approx 3.8442$ *are absolute constants.*

Proof: This proof is based on large deviation theory for Gaussian matrices. Note that by the definition of the reshaping permutation operator $\mathcal{R}(\cdot)$, we have:

$$\mathbf{\Delta}_n = \frac{1}{n}\sum_{t=1}^{n} \begin{bmatrix} \text{vec}(\mathbf{z}_t(1)\mathbf{z}_t(1)^T)^T - \mathbb{E}[\text{vec}(\mathbf{z}_t(1)\mathbf{z}_t(1)^T)^T] \\ \vdots \\ \text{vec}(\mathbf{z}_t(p)\mathbf{z}_t(p)^T)^T - \mathbb{E}[\text{vec}(\mathbf{z}_t(p)\mathbf{z}_t(p)^T)^T] \end{bmatrix}$$

where $\mathbf{z}_t(i) = [\mathbf{z}_t]_{(i-1)q+1:iq}$ is the ith subvector of tth observation $\mathbf{z}_t$. Thus, we can write:

$$\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y} = \frac{1}{n}\sum_{t=1}^{n}\psi_t$$

where

$$\begin{aligned}\psi_t = &\sum_{i,j=1}^{p}\sum_{k,l=1}^{q}\mathbf{X}_{i,j}\mathbf{Y}_{k,l} \\ &\times\left([\mathbf{z}_t]_{(i-1)q+k}[\mathbf{z}_t]_{(j-1)q+l} - \mathbb{E}[[\mathbf{z}_t]_{(i-1)q+k}[\mathbf{z}_t]_{(j-1)q+l}]\right) \end{aligned} \tag{32}$$

and $\mathbf{X} \in \mathbb{R}^{p\times p}$ and $\mathbf{Y} \in \mathbb{R}^{q\times q}$ are reshaped versions of $\mathbf{x}$ and $\mathbf{y}$. Defining $\mathbf{M} = \mathbf{X}\otimes\mathbf{Y}$, we can write (32) as:

$$\psi_t = \mathbf{z}_t^T\mathbf{M}\mathbf{z}_t - \mathbb{E}[\mathbf{z}_t^T\mathbf{M}\mathbf{z}_t]$$

The statistic (32) has the form of Gaussian chaos of order 2. Many of the random variables involved in the summation (32) are correlated, which makes the analysis difficult. To simplify the concentration of measure derivation, using the joint Gaussian property of the data, we note that the stochastic equivalent of $\mathbf{z}_t^T\mathbf{M}\mathbf{z}_t$ is $\boldsymbol{\beta}_t^T\tilde{\mathbf{M}}\boldsymbol{\beta}_t$, where $\tilde{\mathbf{M}} = \mathbf{\Sigma}_0^{1/2}\mathbf{M}\mathbf{\Sigma}_0^{1/2}$, where $\boldsymbol{\beta}_t \sim N(\mathbf{0}, \mathbf{I}_{pq})$ is a random vector with i.i.d. standard normal components. By this decoupling argument, we have:

$$\begin{aligned}\mathbb{E}|\psi_t|^2 &= \mathbb{E}\left|\boldsymbol{\beta}_t^T\tilde{\mathbf{M}}\boldsymbol{\beta}_t - \mathbb{E}[\boldsymbol{\beta}_t^T\tilde{\mathbf{M}}\boldsymbol{\beta}_t]\right|^2 \\ &= \mathbb{E}\left|\sum_{i_1\neq i_2}[\boldsymbol{\beta}_t]_{i_1}[\boldsymbol{\beta}_t]_{i_2}\tilde{\mathbf{M}}_{i_1,i_2} + \sum_{i_1=1}^{d}([\boldsymbol{\beta}_t]_{i_1}^2 - 1)\tilde{\mathbf{M}}_{i_1,i_1}\right|^2 \\ &= \sum_{i_1\neq i_2}\sum_{i_1'\neq i_2'}\mathbb{E}[[\boldsymbol{\beta}_t]_{i_1}[\boldsymbol{\beta}_t]_{i_2}[\boldsymbol{\beta}_t]_{i_1'}[\boldsymbol{\beta}_t]_{i_2'}]\tilde{\mathbf{M}}_{i_1,i_2}\tilde{\mathbf{M}}_{i_1',i_2'} \\ &\quad + \sum_{i_1}\sum_{i_1'}\mathbb{E}[([\boldsymbol{\beta}_t]_{i_1}^2 - 1)([\boldsymbol{\beta}_t]_{i_1'}^2 - 1)]\tilde{\mathbf{M}}_{i_1,i_1}\tilde{\mathbf{M}}_{i_1',i_1'} \\ &= \sum_{i_1\neq i_2}\tilde{\mathbf{M}}_{i_1,i_2}^2 + 2\sum_{i_1}\tilde{\mathbf{M}}_{i_1,i_1}^2 \\ &= \|\tilde{\mathbf{M}}\|_F^2 + \|\text{diag}(\tilde{\mathbf{M}})\|_F^2 \\ &\leq 2\|\tilde{\mathbf{M}}\|_F^2 \leq 2\|\mathbf{\Sigma}_0\|_2^2\|\mathbf{M}\|_F^2 = 2\|\mathbf{\Sigma}_0\|_2^2\end{aligned}$$

where in the last step we used $\|\mathbf{M}\|_F = \|\mathbf{X}\|_F \|\mathbf{Y}\|_F = 1$.

It can be shown (see Appendix A in [45]) that for all $m \geq 2$:

$$\mathbb{E}|\psi_t|^m \leq m! W^{m-2} v_t/2 \tag{33}$$

where

$$W = e\sqrt{\mathbb{E}|\psi_t|^2} \leq e\sqrt{2}\|\mathbf{\Sigma}_0\|_2$$
$$v_t = \frac{2e}{\sqrt{6\pi}}\mathbb{E}|\psi_t|^2 \leq \frac{4e}{\sqrt{6\pi}}\|\mathbf{\Sigma}_0\|_2^2$$

From Bernstein's inequality (see Thm. 1.1 in [45]), we obtain:

$$\mathbb{P}(|\frac{1}{n}\sum_{t=1}^{n}\psi_t| \geq \tau) \leq 2\exp\left(\frac{-n^2\tau^2/2}{nv_1 + Wn\tau}\right)$$
$$\leq 2\exp\left(\frac{-n\tau^2/2}{C_1\|\mathbf{\Sigma}_0\|_2^2 + C_2\|\mathbf{\Sigma}_0\|_2\tau}\right)$$

This concludes the proof. ■

APPENDIX D
PROOF OF THEOREM 3

Proof: Let $\mathcal{N}(\mathcal{S}^{d'-1}, \epsilon')$ denote an ϵ'-net on the sphere $\mathcal{S}^{d'-1}$. Let $\mathbf{x}_1 \in \mathcal{S}^{p^2-1}$ and $\mathbf{y}_1 \in \mathcal{S}^{q^2-1}$ be such that $|<\mathbf{x}_1, \mathbf{\Delta}_n\mathbf{y}_1>| = \|\mathbf{\Delta}_n\|_2$. Let $\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq \epsilon'$ and $\|\mathbf{y}_1 - \mathbf{y}_2\|_2 \leq \epsilon'$. By the Cauchy-Schwarz inequality, for $\mathbf{x}_2 \in \mathcal{N}(\mathcal{S}^{p^2-1}, \epsilon'), \mathbf{y}_2 \in \mathcal{N}(\mathcal{S}^{q^2-1}, \epsilon')$:

$$|\mathbf{x}_1^T\mathbf{\Delta}_n\mathbf{y}_1| - |\mathbf{x}_2^T\mathbf{\Delta}_n\mathbf{y}_2| \leq |\mathbf{x}_1^T\mathbf{\Delta}_n\mathbf{y}_1 - \mathbf{x}_2^T\mathbf{\Delta}_n\mathbf{y}_2|$$
$$= |<\mathbf{x}_1, \mathbf{\Delta}_n(\mathbf{y}_1 - \mathbf{y}_2)> + <\mathbf{x}_1 - \mathbf{x}_2, \mathbf{\Delta}_n\mathbf{y}_2>|$$
$$\leq 2\epsilon'\|\mathbf{\Delta}_n\|_2$$

This implies:

$$\|\mathbf{\Delta}_n\|_2 \leq (1-2\epsilon')^{-1} \max_{\mathbf{x}\in\mathcal{N}(\mathcal{S}^{p^2-1},\epsilon'), \mathbf{y}\in\mathcal{N}(\mathcal{S}^{q^2-1},\epsilon')} |\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}| \tag{34}$$

From Lemma 5.2 in [46], we have the bound on the cardinality of the ϵ'-net on the unit Euclidean sphere $S^{d'-1}$:

$$\text{card}(\mathcal{N}(\mathcal{S}^{d'-1}, \epsilon')) \leq \left(1 + \frac{2}{\epsilon'}\right)^{d'} \tag{35}$$

From (34), (35) and the union bound:

$$\begin{aligned}
&\mathbb{P}(\|\mathbf{\Delta}_n\|_2 \geq \epsilon) \\
&\quad \leq \mathbb{P}\left(\max_{\mathbf{x}\in\mathcal{N}(\mathcal{S}^{p^2-1},\epsilon'),\mathbf{y}\in\mathcal{N}(\mathcal{S}^{q^2-1},\epsilon')} |\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon')\right) \\
&\quad \leq \mathbb{P}\left(\bigcup_{\mathbf{x}\in\mathcal{N}(\mathcal{S}^{p^2-1},\epsilon'),\mathbf{y}\in\mathcal{N}(\mathcal{S}^{q^2-1},\epsilon')} |\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon')\right) \\
&\quad \leq \text{card}(\mathcal{N}(\mathcal{S}^{p^2-1},\epsilon'))\text{card}(\mathcal{N}(\mathcal{S}^{q^2-1},\epsilon')) \\
&\qquad \times \max_{\mathbf{x}\in\mathcal{N}(\mathcal{S}^{p^2-1},\epsilon'),\mathbf{y}\in\mathcal{N}(\mathcal{S}^{q^2-1},\epsilon')} \mathbb{P}(|\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon')) \\
&\quad \leq \left(1+\frac{2}{\epsilon'}\right)^{p^2+q^2} \mathbb{P}\left(|\mathbf{x}^T\mathbf{\Delta}_n\mathbf{y}| \geq \epsilon(1-2\epsilon')\right)
\end{aligned}$$

Using Lemma 2, we further obtain:

$$\begin{aligned}
&\mathbb{P}(\|\mathbf{\Delta}_n\|_2 \geq \epsilon) \\
&\quad \leq 2\left(1+\frac{2}{\epsilon'}\right)^{p^2+q^2} \exp\left(\frac{-n\epsilon^2(1-2\epsilon')^2/2}{C_1\|\mathbf{\Sigma}_0\|_2^2 + C_2\|\mathbf{\Sigma}_0\|_2\epsilon(1-2\epsilon')}\right)
\end{aligned} \tag{36}$$

We finish the proof by considering the two separate regimes. First, let us consider the Gaussian tail regime which occurs when $n > (\frac{tC_2}{C_1})^2(p^2+q^2+\log M)$ and choose:

$$\epsilon = \frac{t\|\mathbf{\Sigma}_0\|_2}{1-2\epsilon'}\sqrt{\frac{p^2+q^2+\log M}{n}}$$

For this regime, the bound (36) can be relaxed to:

$$\begin{aligned}
&\mathbb{P}(\|\mathbf{\Delta}_n\|_2 \geq \epsilon) \\
&\quad \leq 2\left(1+\frac{2}{\epsilon'}\right)^{p^2+q^2} \exp\left(\frac{-n\epsilon^2(1-2\epsilon')^2/2}{2C_1\|\mathbf{\Sigma}_0\|_2^2}\right)
\end{aligned} \tag{37}$$

Then, from (37), we have:

$$\begin{aligned}
&\mathbb{P}\left(\|\mathbf{\Delta}_n\|_2 \geq \frac{t\|\mathbf{\Sigma}_0\|_2}{1-2\epsilon'}\sqrt{\frac{p^2+q^2+\log M}{n}}\right) \\
&\quad \leq 2\left(1+\frac{2}{\epsilon'}\right)^{p^2+q^2} \exp\left(\frac{-t^2(p^2+q^2+\log M)}{4C_1}\right) \\
&\quad \leq 2\left(\left(1+\frac{2}{\epsilon'}\right)e^{-t^2/(4C_1)}\right)^{p^2+q^2} M^{-t^2/(4C_1)} \\
&\quad \leq 2M^{-t^2/(4C_1)}
\end{aligned}$$

This concludes the bound for the first regime. Second, the exponential tail regime follows by similar arguments. Assuming $n \leq \frac{tC_2}{C_1}(p^2+q^2+\log M)$, we obtain from (36):

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{\Delta}_n\|_2 \geq \frac{t\|\boldsymbol{\Sigma}_0\|_2}{1-2\epsilon'}\frac{p^2+q^2+\log M}{n}\right) \\
&\leq 2\left(1+\frac{2}{\epsilon'}\right)^{p^2+q^2}\exp\left(\frac{-t(p^2+q^2+\log M)}{4C_2}\right) \\
&\leq 2\left(\left(1+\frac{2}{\epsilon'}\right)e^{-t/(4C_2)}\right)^{p^2+q^2} M^{-t/(4C_2)} \\
&\leq 2M^{-t/(4C_2)}
\end{aligned}$$

where we used the assumption $t \geq 4C_2\ln(1+\frac{2}{\epsilon'})$. The proof is complete by combining both regimes and taking $C_0 > 0$ large enough [3] and noting that $t > 1$. ■

APPENDIX E
PROOF OF THEOREM 4

Proof: Define the event

$$E_r = \left\{\|\hat{\mathbf{R}}_n^{\lambda} - \mathbf{R}_0\|_F^2 > \inf_{\mathbf{R}:\text{rank}(\mathbf{R})\leq r}\|\mathbf{R}-\mathbf{R}_0\|_F^2 + \frac{(1+\sqrt{2})^2}{4}\lambda_n^2 r\right\}$$

where λ_n is chosen as in the statement of the theorem.

Theorem 2 implies that on the event $\lambda \geq 2\|\boldsymbol{\Delta}_n\|_2$, with probability 1, we have for any $1 \leq r \leq r_0$:

$$\|\hat{\mathbf{R}}_n^{\lambda} - \mathbf{R}_0\|_F^2 \leq \inf_{\mathbf{R}:\text{rank}(\mathbf{R})\leq r}\|\mathbf{R}-\mathbf{R}_0\|_F^2 + \frac{(1+\sqrt{2})^2}{4}\lambda^2 r$$

Using this and Theorem 3, we obtain:

$$\begin{aligned}
\mathbb{P}(E_r) &= \mathbb{P}(E_r \cap \{\lambda_n \geq 2\|\boldsymbol{\Delta}_n\|_2\}) + \mathbb{P}(E_r \cap \{\lambda_n < 2\|\boldsymbol{\Delta}_n\|_2\}) \\
&\leq 0 + \mathbb{P}(\lambda_n < 2\|\boldsymbol{\Delta}_n\|_2) \\
&= \mathbb{P}\Big(\|\boldsymbol{\Delta}_n\|_2 > \frac{C_0 t}{1-2\epsilon'} \\
&\quad \times \max\left\{\frac{p^2+q^2+\log M}{n}, \sqrt{\frac{p^2+q^2+\log M}{n}}\right\}\Big) \\
&\leq 2M^{-t/4C}
\end{aligned}$$

This concludes the proof.

■

[3] Note that this constant depends on the constants $t, C_1, C_2 > 0$.

APPENDIX F
PROOF OF LEMMA 1

Proof: From the min-max theorem of Courant-Fischer-Weyl [47]:

$$\begin{aligned}\sigma_{k+1}^2(\mathbf{R}) &= \lambda_{k+1}(\mathbf{R}\mathbf{R}^T) \\ &= \min_{V:\dim(V^\perp)\leq k} \max_{\|\mathbf{v}\|_2=1, \mathbf{v}\in V} < \mathbf{R}\mathbf{R}^T\mathbf{v}, \mathbf{v} >\end{aligned}$$

Define the set $\mathcal{V}_k = \{\mathbf{v} \in \mathbb{R}^{p^2} : \|\mathbf{v}\|_2 = 1, \mathbf{v} \perp \text{Col}(\mathbf{R}\mathbf{P}_k\mathbf{R}^T)\} \subset S^{p^2-1}$. Choosing $V = \text{Col}(\mathbf{R}\mathbf{P}_k\mathbf{R}^T)^\perp$, we have the upper bound:

$$\sigma_{k+1}^2(\mathbf{R}) \leq \max_{\mathbf{v}\in\mathcal{V}_k} < \mathbf{R}\mathbf{R}^T\mathbf{v}, \mathbf{v} >$$

Using the definition of $\mathcal{V}_k$ and the orthogonality principle, we have:

$$\begin{aligned}< \mathbf{R}\mathbf{R}^T\mathbf{v}, \mathbf{v} > &=< \mathbf{R}(\mathbf{I} - \mathbf{P}_k)\mathbf{R}^T\mathbf{v}, \mathbf{v} > \\ &=< (\mathbf{I} - \mathbf{P}_k)\mathbf{R}^T\mathbf{v}, \mathbf{R}^T\mathbf{v} > \\ &=< (\mathbf{I} - \mathbf{P}_k)\mathbf{R}^T\mathbf{v}, (\mathbf{I} - \mathbf{P}_k)\mathbf{R}^T\mathbf{v} > \\ &= \|(\mathbf{I} - \mathbf{P}_k)\mathbf{R}^T\mathbf{v}\|_2^2\end{aligned}$$

Using this equality and the definition of the spectral norm [47]:

$$\begin{aligned}\sigma_{k+1}^2(\mathbf{R}) &\leq \max_{\mathbf{v}\in\mathcal{V}_k} \|(\mathbf{I} - \mathbf{P}_k)\mathbf{R}^T\mathbf{v}\|_2^2 \\ &\leq \max_{\mathbf{v}\in S^{p^2-1}} \|(\mathbf{I} - \mathbf{P}_k)\mathbf{R}^T\mathbf{v}\|_2^2 \\ &= \|(\mathbf{I} - \mathbf{P}_k)\mathbf{R}^T\|_2^2\end{aligned}$$

Equality follows when choosing $\mathbf{P}_k = \mathbf{V}_k\mathbf{V}_k^T$. This is seen by writing $\mathbf{I} = \mathbf{V}\mathbf{V}^T$ and using the definition of the spectral norm and the sorting of the singular values. The proof is complete.

■

APPENDIX G
PROOF OF THEOREM 5

Proof: Note that $(\lambda, \mathbf{u})$ is an eigenvalue-eigenvector pair of the square symmetric matrix $\mathbf{R}_0^T\mathbf{R}_0$ if:

$$\sum_{i,j} \text{vec}(\mathbf{\Sigma}_0(i,j)) < \mathbf{u}, \text{vec}(\mathbf{\Sigma}_0(i,j)) >= \lambda\mathbf{u} \tag{38}$$

So for $\lambda > 0$, the eigenvector $\mathbf{u}$ must lie in the span of the vectorized submatrices $\{\text{vec}(\mathbf{\Sigma}_0(i,j))\}_{i,j}$. Motivated by this result, we use the Gram-Schmidt procedure to construct a basis that incrementally

spans more and more of the subspace span($\{\text{vec}(\mathbf{\Sigma}_0(i,j))\}_{i,j}$). For the special case of the block-Toeplitz matrix, we have:

$$\text{span}(\{\text{vec}(\mathbf{\Sigma}_0(i,j))\}_{i,j}) = \text{span}(\{\text{vec}(\mathbf{\Sigma}(\tau))\}_{\tau=-N}^{N})$$

By stationarity, note that $\mathbf{\Sigma}(-\tau) = \mathbf{\Sigma}(\tau)^T$.

For simplicity, consider the case $k = 2k' + 1$ for some $k' \geq 0$. From Lemma 1, we are free to choose an orthonormal basis set $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ and form the projection matrix $\mathbf{P}_k = \mathbf{V}_k \mathbf{V}_k^T$, where the columns of $\mathbf{V}_k$ are the vectors $\{\mathbf{v}_j\}$. We form the orthonormal basis using the Gram-Schmidt procedure [47]:

$$\begin{aligned}
\tilde{\mathbf{v}}_0 &= \text{vec}(\mathbf{\Sigma}(0)), \\
\mathbf{v}_0 &= \frac{\tilde{\mathbf{v}}_0}{\|\tilde{\mathbf{v}}_0\|_2} \\
\tilde{\mathbf{v}}_1 &= \text{vec}(\mathbf{\Sigma}(1)) - \frac{< \text{vec}(\mathbf{\Sigma}(1)), \tilde{\mathbf{v}}_0 >}{\|\tilde{\mathbf{v}}_0\|_2^2} \tilde{\mathbf{v}}_0, \\
\mathbf{v}_1 &= \frac{\tilde{\mathbf{v}}_1}{\|\tilde{\mathbf{v}}_1\|_2} \\
\tilde{\mathbf{v}}_{-1} &= \text{vec}(\mathbf{\Sigma}(-1)) - \frac{< \text{vec}(\mathbf{\Sigma}(-1)), \tilde{\mathbf{v}}_0 >}{\|\tilde{\mathbf{v}}_0\|_2^2} \tilde{\mathbf{v}}_0 \\
&\quad - \frac{< \text{vec}(\mathbf{\Sigma}(-1)), \tilde{\mathbf{v}}_1 >}{\|\tilde{\mathbf{v}}_1\|_2^2} \tilde{\mathbf{v}}_1, \\
\mathbf{v}_{-1} &= \frac{\tilde{\mathbf{v}}_{-1}}{\|\tilde{\mathbf{v}}_{-1}\|_2} \\
&etc.
\end{aligned}$$

With this choice of orthonormal basis, it follows that for every $k = 2k' + 1$, we have the orthogonal projector:

$$\mathbf{P}_k = \mathbf{v}_0 \mathbf{v}_0^T + \sum_{l=1}^{k'} (\mathbf{v}_l \mathbf{v}_l^T + \mathbf{v}_{-l} \mathbf{v}_{-l}^T)$$

This corresponds to a variant of a sequence of Householder transformations [47]. Using Lemma 1:

$$\begin{aligned}
\sigma_{k+1}^2(\mathbf{R}_0) &\leq \|\mathbf{R}_0(\mathbf{I}-\mathbf{P}_k)\|_2^2 \\
&\leq \|\mathbf{R}_0 - \mathbf{R}_0\mathbf{P}_k\|_F^2 \\
&\leq p\sum_{l=k'+1}^{N} \|\mathbf{\Sigma}(l)\|_F^2 + \|\mathbf{\Sigma}(-l)\|_F^2 \qquad (39)\\
&\leq 2C'pq\sum_{l=k'+1}^{N} u^{2l} \\
&\leq 2C'pq\frac{u^{2k'+2}}{1-u^2} \\
&\leq 2C'pq\frac{u^{k}}{1-u^2}
\end{aligned}$$

where we used Lemma 3 to obtain (39). To finish the proof, using the bound above and (12):

$$\begin{aligned}
\inf_{\mathbf{R}:\text{rank}(\mathbf{R})\leq r} \|\mathbf{R}-\mathbf{R}_0\|_F^2 &= \sum_{k=r}^{r_0-1} \sigma_{k+1}^2(\mathbf{R}_0) \\
&\leq \frac{2C'pq}{1-u^2}\sum_{k=r}^{r_0-1} u^k \\
&\leq 2C'pq\frac{u^r}{(1-u)^2}
\end{aligned}$$

The proof is complete.

■

APPENDIX H
LEMMA 3

Lemma 3. *Consider the notation and setting of proof of Thm. 5. Then, for the projection matrix* $\mathbf{P}_k$ *chosen, we have for* $k = 2k'+1, k' \geq 1$*:*

$$\sigma_{k+1}^2(\mathbf{R}_0) \leq \|\mathbf{R}_0 - \mathbf{R}_0\mathbf{P}_k\|_F^2 \leq p\sum_{l=k'+1}^{N} \|\mathbf{\Sigma}(l)\|_F^2 + \|\mathbf{\Sigma}(-l)\|_F^2$$

Proof: To illustrate the row-subtraction technique, we consider the simplified scenario $k' = 1$. The proof can be easily generalized to all $k' \geq 1$. Without loss of generality, we write the permuted covariance

$$\mathbf{\Sigma}_0 = \begin{bmatrix} \mathbf{\Sigma}(0) & \mathbf{\Sigma}(1) \\ \mathbf{\Sigma}(-1) & \mathbf{\Sigma}(0) \end{bmatrix} \qquad (40)$$

as:

$$\mathbf{R}_0 = \mathbf{R}(\mathbf{\Sigma}_0) = \begin{bmatrix} \text{vec}(\mathbf{\Sigma}(0))^T \\ \text{vec}(\mathbf{\Sigma}(1))^T \\ \text{vec}(\mathbf{\Sigma}(-1))^T \\ \text{vec}(\mathbf{\Sigma}(0))^T \end{bmatrix}$$

Using the Gram-Schmidt submatrix basis construction of the proof of Thm. 5, the sequence of projection matrices can be written as:

$$\mathbf{P}_1 = \mathbf{v}_0\mathbf{v}_0^T$$

$$\mathbf{P}_2 = \mathbf{v}_0\mathbf{v}_0^T + \mathbf{v}_1\mathbf{v}_1^T$$

$$\mathbf{P}_3 = \mathbf{v}_0\mathbf{v}_0^T + \mathbf{v}_1\mathbf{v}_1^T + \mathbf{v}_{-1}\mathbf{v}_{-1}^T$$

where $\mathbf{v}_i$ is the orthonormal basis constructed in the proof of Thm. 5. The singular value bound $\sigma_1^2(\mathbf{R}_0) \leq \|\mathbf{R}_0\|_F^2 = 2\|\mathbf{\Sigma}(0)\|_F^2 + \|\mathbf{\Sigma}(1)\|_F^2 + \|\mathbf{\Sigma}(-1)\|_F^2$ is trivial [47].

For the second singular value, we want to prove the bound:

$$\sigma_2^2(\mathbf{R}_0) \leq \|\mathbf{\Sigma}(1)\|_F^2 + \|\mathbf{\Sigma}(-1)\|_F^2 \tag{41}$$

To show this, we use the variational bound of Lemma 1:

$$\begin{aligned}
\sigma_2^2(\mathbf{R}_0) &\leq \|\mathbf{R}_0 - \mathbf{R}_0\mathbf{P}_1\|_F^2 \\
&= \left\| \begin{bmatrix} \text{vec}(\mathbf{\Sigma}(0))^T - <\text{vec}(\mathbf{\Sigma}(0)), \mathbf{v}_0>\mathbf{v}_0^T \\ \text{vec}(\mathbf{\Sigma}(1))^T - <\text{vec}(\mathbf{\Sigma}(1)), \mathbf{v}_0>\mathbf{v}_0^T \\ \text{vec}(\mathbf{\Sigma}(-1))^T - <\text{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_0>\mathbf{v}_0^T \\ \text{vec}(\mathbf{\Sigma}(0))^T - <\text{vec}(\mathbf{\Sigma}(0)), \mathbf{v}_0>\mathbf{v}_0^T \end{bmatrix} \right\|_F^2 \\
&= \left\| \begin{bmatrix} \mathbf{0}^T \\ \text{vec}(\mathbf{\Sigma}(1))^T - <\text{vec}(\mathbf{\Sigma}(1)), \mathbf{v}_0>\mathbf{v}_0^T \\ \text{vec}(\mathbf{\Sigma}(-1))^T - <\text{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_0>\mathbf{v}_0^T \\ \mathbf{0}^T \end{bmatrix} \right\|_F^2 \\
&= \|\text{vec}(\mathbf{\Sigma}(1)) - <\text{vec}(\mathbf{\Sigma}(1)), \mathbf{v}_0>\mathbf{v}_0\|_2^2 \\
&\quad + \|\text{vec}(\mathbf{\Sigma}(-1)) - <\text{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_0>\mathbf{v}_0\|_2^2 \\
&\leq \|\mathbf{\Sigma}(1)\|_F^2 + \|\mathbf{\Sigma}(-1)\|_F^2
\end{aligned}$$

where in the last step, we used the Pythagorean principle from least-squares theory [48]-i.e.$\|\mathbf{A} - \frac{<\mathbf{A},\mathbf{B}>}{\|\mathbf{B}\|_F^2}\mathbf{B}\|_F^2 \leq \|\mathbf{A}\|_F^2$ for any matrices $\mathbf{A}, \mathbf{B}$ of the same order. Next, we want to show

$$\sigma_3^2(\mathbf{R}_0) \leq \|\mathbf{\Sigma}(-1)\|_F^2 \tag{42}$$

Define $\gamma(j) = \text{vec}(\mathbf{\Sigma}(j)) - < \text{vec}(\mathbf{\Sigma}(j)), \mathbf{v}_0 > \mathbf{v}_0$. Using similar bounds and the above, after some algebra:

$$\begin{aligned}
\sigma_3^2(\mathbf{R}_0) &\leq \|\mathbf{R}_0 - \mathbf{R}_0\mathbf{P}_2\|_F^2 \\
&= \left\| \begin{bmatrix} \mathbf{0}^T \\ \gamma(1)^T - < \text{vec}(\mathbf{\Sigma}(1)), \mathbf{v}_1 > \mathbf{v}_1^T \\ \gamma(-1)^T - < \text{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_1 > \mathbf{v}_1^T \\ \mathbf{0}^T \end{bmatrix} \right\|_F^2 \\
&= \|\text{vec}(\mathbf{\Sigma}(-1))^T - < \text{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_0 > \mathbf{v}_0^T \\
&\quad - < \text{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_1 > \mathbf{v}_1^T\|_2^2 \\
&= \|\text{vec}(\mathbf{\Sigma}(-1))^T\|_2^2 - | < \text{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_0 > |^2 \\
&\quad - | < \text{vec}(\mathbf{\Sigma}(-1)), \mathbf{v}_1 > |^2 \\
&\leq \|\mathbf{\Sigma}(-1)\|_F^2
\end{aligned}$$

where we observed that $\gamma(1) =< \text{vec}(\mathbf{\Sigma}(1)), \mathbf{v}_1 > \mathbf{v}_1$ and used the Pythagorean principle again.

Using $\mathbf{P}_3$ and similar bounds, it follows that $\sigma_4^2(\mathbf{R}_0) = 0$, which makes sense since the separation rank of (40) is at most 3. Generalizing to $k' \geq 1$ and noting that $\|\mathbf{\Sigma}_0\|_F^2 = p\|\mathbf{\Sigma}(0)\|_F^2 + \sum_{l=1}^{p-1}(p-l)\|\mathbf{\Sigma}(l)\|_F^2 + \|\mathbf{\Sigma}(l)\|_F^2 \leq p\|\mathbf{\Sigma}(0)\|_F^2 + p\sum_{l=1}^{p-1}\|\mathbf{\Sigma}(l)\|_F^2 + \|\mathbf{\Sigma}(-l)\|_F^2$, we conclude the proof. ■

REFERENCES

[1] K. Lounici, "High-dimensional covariance matrix estimation with missing observations," *arXiv:1201.2577v5*, May 2012.

[2] T. Tsiligkaridis, A. Hero, and S. Zhou, "On Convergence of Kronecker Graphical Lasso Algorithms," *to appear in IEEE Transactions on Signal Processing*, 2013.

[3] ——, "Convergence properties of kronecker graphical lasso algorithms," *arXiv:1204.0585*, July 2012.

[4] J. Bai and S. Shi, "Estimating high dimensional covariance matrices and its applications," *Annals of Economics and Finance*, vol. 12, no. 2, pp. 199–215, 2011.

[5] J. Xie and P. M. Bentler, "Covariance structure models for gene expression microarray data," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 10, no. 4, pp. 556–582, 2003.

[6] A. Hero and B. Rajaratnam, "Hub discovery in partial correlation graphs," *IEEE Transactions on Information Theory*, vol. 58, no. 9, pp. 6064–6078, September 2012.

[7] G. Derado, F. D. Bowman, and C. D. Kilts, "Modeling the spatial and temporal dependence in fmri data," *Biometrics*, vol. 66, no. 3, pp. 949–957, September 2010.

[8] Y. Zhang and J. Schneider, "Learning multiple tasks with a sparse matrix-normal penalty," *Advances in Neural Information Processing Systems*, vol. 23, pp. 2550–2558, 2010.

[9] G. I. Allen and R. Tibshirani, "Transposable regularized covariance models with an application to missing data imputation," *The Annals of Applied Statistics*, vol. 4, no. 2, pp. 764–790, 2010.

[10] M. Yuan and Y. Lin, "Model selection and estimation in the gaussian graphical model," *Biometrika*, vol. 94, pp. 19–35, 2007.

[11] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *Journal of Machine Learning Research*, vol. 9, pp. 485–516, March 2008.

[12] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Advances in Neural Information Processing Systems*, 2008.

[13] A. Rothman, P. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.

[14] J. Fan, Y. Fan, and J. Lv, "High dimensional covariance matrix estimation using a factor model," *Journal of Econometrics*, vol. 147, no. 1, pp. 1348–1360, 2008.

[15] G. Fitzmaurice, N. Laird, and J. Ware, *Applied longitudinal analysis*. Wiley-Interscience, 2004.

[16] I. Johnstone and A. Lu, "On consistency and sparsity for principal component analysis in high dimensions," *Journal of the American Statstical Association*, vol. 104, no. 486, pp. 682–693, 2009.

[17] P. Dutilleul, "The mle algorithm for the matrix normal distribution," *J. Statist. Comput. Simul.*, vol. 64, pp. 105–123, 1999.

[18] K. Werner, M. Jansson, and P. Stoica, "On estimation of covariance matrices with Kronecker product structure," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, February 2008.

[19] A. Dawid, "Some matrix-variate distribution theory: notational considerations and a bayesian application," *Biometrika*, vol. 68, pp. 265–274, 1981.

[20] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*. Chapman Hill, 1999.

[21] K. Werner and M. Jansson, "Estimation of kronecker structured channel covariances using training data," in *Proceedings of EUSIPCO*, 2007.

[22] N. Cressie, *Statistics for Spatial Data*. Wiley, New York, 1993.

[23] J. Yin and H. Li, "Model selection and estimation in the matrix normal graphical model," *Journal of Multivariate Analysis*, vol. 107, pp. 119–140, 2012.

[24] E. Bonilla, K. M. Chai, and C. Williams, "Multi-task gaussian process prediction," *Advances in Neural Information Processing Systems*, pp. 153–160, 2008.

[25] K. Yu, J. Lafferty, S. Zhu, and Y. Gong, "Large-scale collaborative prediction using a nonparametric random effects model," *ICML*, pp. 1185–1192, 2009.

[26] G. Beylkin and M. J. Mohlenkamp, "Algorithms for numerical analysis in high dimensions," *SIAM Journal on Scientific Computing*, vol. 26, no. 6, pp. 2133–2159, 2005.

[27] J. C. de Munck, H. M. Huizenga, L. J. Waldorp, and R. M. Heethaar, "Estimating stationary dipoles from meg/eeg data contaminated with spatially and temporally correlated background noise," *IEEE Transactions on Signal Processing*, vol. 50, no. 7, July 2002.

[28] J. C. de Munck, F. Bijma, P. Gaura, C. A. Sieluzycki, M. I. Branco, and R. M. Heethaar, "A maximum-likelihood estimator

for trial-to-trial variations in noisy meg/eeg data sets," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 12, 2004.

[29] F. Bijma, J. de Munck, and R. Heethaar, "The spatiotemporal meg covariance matrix modeled as a sum of kronecker products," *NeuroImage*, vol. 27, pp. 402–415, 2005.

[30] S. C. Jun, S. M. Plis, D. M. Ranken, and D. M. Schmidt, "Spatiotemporal noise covariance estimation from limited empirical magnetoencephalographic data," *Physics in Medicine and Biology*, vol. 51, pp. 5549–5564, 2006.

[31] A. Rucci, S. Tebaldini, and F. Rocca, "Skp-shrinkage estimator for sar multi-baselines applications," in *Proceedings of IEEE Radar Conference*, 2010.

[32] C. V. Loan and N. Pitsianis, "Approximation with kronecker products," in *Linear Algebra for Large Scale and Real Time Applications*. Kluwer Publications, 1993, pp. 293–314.

[33] C. Leng and C. Y. Tang, "Sparse matrix graphical models," *Journal of the American Statistical Association*, vol. 107, pp. 1187–1200, October 2012.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[35] J.-F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal of Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[36] J.-F. Cai and S. Osher, "Fast singular value thresholding without singular value decomposition," UCLA, Tech. Rep., 2010.

[37] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, May 2011.

[38] J. Haslett and A. E. Raftery, "Space-time modeling with long-memory dependence: assessing ireland's wind power resource," *Applied Statistics*, vol. 38, no. 1, pp. 1–50, 1989.

[39] T. Gneiting, "Nonseparable, stationary covariance functions for space-time data," *Journal of the American Statistical Association (JASA)*, vol. 97, no. 458, pp. 590–600, 2002.

[40] X. de Luna and M. Genton, "Predictive spatio-temporal models for spatially sparse environmental data," *Statistica Sinica*, vol. 15, pp. 547–568, 2005.

[41] M. Stein, "Space-time covariance functions," *Journal of the American Statistical Association (JASA)*, vol. 100, pp. 310–321, 2005.

[42] Y. Chen, A. Wiesel, and A. Hero, "Robust shrinkage estimation of high dimensional covariance matrices," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4097–4107, September 2011.

[43] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W.Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph, "The ncep/ncar 40-year reanalysis project," *Bulletin of the American Meteorological Society*, vol. 77, no. 3, p. 437471, 1996.

[44] G. A. Watson, "Characterization of the subdifferential of some matrix norms," *Linear Algebra and Applications*, vol. 170, pp. 33–45, 1992.

[45] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Transactions on Information Theory*, May 2008.

[46] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv:1011.3027v7*, November 2011.

[47] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 1st ed. Cambridge University Press, 1990.

[48] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*. Mc Graw Hill, 2002.